Improvement of methods for the structural characterisation of drug metabolites based on collisional cross sections

Dmytro Ivashchenko

Doctor’s Degree in Chemistry

Department of Chemistry and Biochemistry

2022

Supervisor

Alexandre Lopes de Magalhães, University of Porto, Faculty of Sciences, Department of Chemistry and Biochemistry

Co-supervisor

Inés Corral Pérez, Autonomous University of Madrid, Faculty of Sciences, Chemistry Department

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

UAM
Universidad Autónoma
de Madrid

PhD

In memory of my father

To my mother with love and eternal appreciation

Sworn Statement

I, Dmytro Ivashchenko, born in Voronizh, Sumy region, Ukraine, resident in Ukraine, phone number +380 967534360, of Ukrainian nationality, bearer of Passport No. FF547775, enrolled in the Doctor's Degree in Chemistry at the Faculty of Sciences of the University of Porto hereby declare, in accordance with the provisions of paragraph a) of Article 14 of the Code of Ethical Conduct of the University of Porto, that the content of this thesis reflects perspectives, research work and my own interpretations at the time of its submission.

By submitting this thesis, I also declare that it contains the results of my own research work and contributions that have not been previously submitted to this or any other institution.

I further declare that all references to other authors fully comply with the rules of attribution and are referenced in the text by citation and identified in the bibliographic references section. This thesis does not include any content whose reproduction is protected by copyright laws.

I am aware that the practice of plagiarism and self-plagiarism constitute a form of academic offense.

Dmytro Ivashchenko

26.07.2022

Acknowledgements

I would like to take this opportunity to express my deepest gratitude to my academic supervisors Prof. Alexandre Lopes de Magalhães and Dr. Inés Corral Pérez and to my industrial supervisor Dr. Jordi Munoz-Muriedas for their enthusiastic encouragement, patient guidance and useful critiques of this research work. Thanks to their valuable and constructive suggestions during the planning and development of this project, I was able to learn new and to improve existing professional skills. My grateful thanks are also extended to the members of the Computational Sciences Department, GlaxoSmithKline and the Global Spectroscopy Department, GlaxoSmithKline, for their help with anything I needed during my secondment and for making my work at GlaxoSmithKline very fruitful. Needless to say I deeply appreciate all the support I received from the people from UCIBIO/REQUIMTE at University of Porto and the Chemistry Department at Autonomous University of Madrid throughout my stays there.

This work would not have been possible without the financial support, provided by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 642294. The project is a part of a broader framework of Innovative Training Network "European Joint Doctorate in Theoretical Chemistry and Computational Modelling" (ITN-EJD-TCCM). I wish to acknowledge enormous amount of work, done by the people involved in this framework, to create and run it, particularly local coordinators at University of Porto and Autonomous University of Madrid and the main coordinator of the network — Prof. Manuel Yáñez.

Last, but not least, I would like to thank my family and friends for their tremendous support and infinite encouragement throughout my PhD. It would have been much harder to make it possible without their essential help.

Thank you. Obrigado. Gracias. Дякую.

Resumo

O metabolismo de um fármaco é um fator muito importante para a variação da sua concentração fisiológica e pode determinar ou modificar a sua atividade farmacológica ou tóxica (Iyanagi, T., Int. Rev. Cytol., 2007, 260). Compreender os processos a que um fármaco está sujeito num organism vivo é, portanto, crucial para estudar e analisar a ação de um fármaco ou dos seus metabolitos como sublinhou Caldwell, J. *et al*, Toxicol. Pathol., 1995, 23 (2). Os metabolitos têm sido identificados por diversas técnicas mas, ultimamente, a espectroscopia de mobilidade iónica (Ion-Mobility Mass Spectrometry — IM-MS) tem vindo a revelar-se uma técnica muito popular para a identificação de metabolitos de pequena dimensão, devido à sua elevada eficiência na análise de amostras de reduzida massa.

A associação desta técnica a metodologias computacionais de cálculo das secções de choque de colisões (collisional cross sections — CCS) tem-se revelado bastante promissora na previsão da estrutura de compostos. Contudo, apesar de nos últimos anos se ter observado um importante desenvolvimentos na correspondente componente experimental, a evolução das metodologias teóricas tem sido mais lenta. Recentemente, Reading, E. *et al*, Anal. Chem., 2016, 88 (4), desenvolveu um protocolo metodológico para cálculo de secções de choque de colisões. A primeira parte deste trabalho aborda a eficiência do protocolo proposto, assim como o seu âmbito de aplicação. Foi também colocada uma atenção especial na reprodutibilidade dos resultados publicados e nas estratégias para melhorar a concordância entre vários conjuntos de resultados teóricos, e entre novos cálculos e resultados experimentais.

A segunda parte da Tese debruça-se sobre o estudo de mecanismos de fragmentação que ocorrem nos ensaios de Espectroscopia de Massa (Mass Spectroscopy — MS). A Ionização por Electrospray (Electro Spray Ionisation — ESI), a Espectrometria de Massa Tandem (Tandem MS) e Dissociação induzida por colisão (Collision Induced Dissociation — CID) constituem ponderosas metodologias experimentais, capazes de proporcionar uma compreensão mais sólida do processo de colisão e dos produtos resultantes (Molina, E. R. *et al*, J. Mass Spectrom., 2015, 50). A abordagem computacional desenvolvida por Hase, W. L. *et al*, Quantum Chem. Progr. Exch. Bull., 1996, 16, and Hase, W. L. *et al*, J. Phys. Chem., 1996, 100(20), é usada para correr simulações por Dinâmica de Colisões (Collision Dynamics Simulations — CDS) de modo a obter trajetórias de reacção. Estas são posteriormemnte utilizadas na análise de fragmentação que permite prever teoricamente a estrutura dos fragmentos, possíveis caminhos de reacção e espectros de massa.

Resumen

El metabolismo del fármaco es un factor determinante esencial para los cambios en la concentración fisiológica del fármaco y puede determinar o modificar su camino toxicológico o farmacológico (Iyanagi, T., Int. Rev. Cytol., 2007, 260). La comprensión de los procesos, que interesan un medicamento en un organismo vivo, es por lo tanto crucial para estudiar y analizar la acción del fármaco o sus metabolitos, según lo reportado por Caldwell, J. *et al*, Toxicol. Pathol., 1995, vol. 23, no. 2. Los metabolitos de los medicamentos se identifican normalmente mediante diversas técnicas, pero últimamente, la espectrometría de masas de movilidad iónica (IM-MS) se ha convertido en una herramienta muy popular para la identificación estructural de moléculas pequeñas (como son los metabolitos de los medicamentos) debido a su alta eficiencia y al requerir baja cantidad de muestra.

La combinación de esta técnica con un enfoque computacional ha demostrado entregar predicciones confiables de identificación de los compuestos investigados al comparar las secciones transversales de colisión (CCS) experimentales y calculadas. Sin embargo, a pesar de los desarrollos valiosos del campo experimental correspondiente en los últimos años, la contraparte teórica ha visto una mejora bastante lenta. Recientemente, Reading, E. *et al*, Anal. Chem., 2016, 88 (4), han desarrollado un protocolo computacional para cálculos de sección transversal colisional. La primera parte de este trabajo aborda el tema de la eficiencia del protocolo propuesto junto con su aplicabilidad a gran escala. Además, se ha prestado especial atención a la reproducibilidad de los resultados publicados y también a las posibles formas de mejorar el acuerdo dentro de diferentes conjuntos de resultados teóricos, así como entre los valores calculados recientemente y los valores experimentales.

La segunda parte de este manuscrito se centra en el estudio de los mecanismos de fragmentación que se producen durante las mediciones de espectrometría de masas (MS). Electro Spray Ionisation (ESI) junto con Tandem MS y Collision Induced Dissociation (CID) construyen un poderoso enfoque experimental, capaz de entregar una comprensión más profunda de un proceso de colisión y sus productos (Molina, ER *et al*, J. Mass Spectrom., 2015, 50). Esto es factible debido a la extensa fragmentación que tiene lugar en los iones activados (metabolitos). Un enfoque computacional correspondiente desarrollado por Hase, W. L. *et al*, Quantum Chem. Progr. Exch. Bull., 1996, 16, y Hase, W. L. *et al*, J. Phys. Chem., 1996, 100.20, se utiliza para ejecutar simulaciones de dinámica de colisión (CDS) para obtener trayectorias reactivas. Estas trayectorias también se utilizan para el análisis de fragmentación que proporciona información sobre la información estructural de los fragmentos y los posibles caminos de reacción y también permite construir un espectro teórico de MS.

Abstract

Drug metabolism is a pivotal determining factor for the changes in physiological drug concentration and can determine or modify its toxicological or pharmacological pathway (Iyanagi, T., Int. Rev. Cytol., 2007, 260). Understanding of processes, involving a drug in a living organism, is therefore crucial to study and analyse the action of the drug or its metabolites, as reported by Caldwell, J. *et al*, Toxicol. Pathol., 1995, vol. 23, no. 2. Drug metabolites are typically identified using various techniques, but lately, Ion-Mobility Mass Spectrometry (IM-MS) has become a widely popular tool for small molecule (which are drug metabolites) structural identification due to its high efficiency and a low amount requirement for samples.

A combination of this technique along with a computational approach has proved to deliver reliable identification predictions of investigated compounds by comparing experimental and calculated collisional cross sections (CCS) of structures. However, even though a corresponding experimental field has made some valuable developments over the last couple of years, its theoretical counterpart has seen a rather slow improvement. Recently, Reading, E. *et al*, Anal. Chem., 2016, 88 (4), have developed a computational protocol for collisional cross section calculations. The first part of this work addresses the issue of efficiency of the proposed protocol along with its large-scale applicability. Additionally, special attention has been paid to the reproducibility of the published results and also to the possible ways of improving the agreement within different sets of theoretical results as well as between newly calculated and experimental values.

The second part of this manuscript focuses on studying fragmentation mechanisms that occur during Mass Spectrometry (MS) measurements. Electro Spray Ionisation (ESI) along with Tandem MS and Collision Induced Dissociation (CID) build up a powerful experimental approach, able to deliver a deeper understanding of a collision process and its products (Molina, E. R. *et al*, J. Mass Spectrom., 2015, 50). It is feasible due to extensive fragmentation that takes place in activated ions (metabolites). A corresponding computational approach developed by Hase, W. L. *et al*, Quantum Chem. Progr. Exch. Bull., 1996, 16, and Hase, W. L. *et al*, J. Phys. Chem., 1996, 100.20, is used to run Collision Dynamics Simulations (CDS) to obtain reactive trajectories. These trajectories are further utilised for fragmentation analysis that gives insights about structural information of the fragments and possible reaction pathways and also allows to build a theoretical MS spectrum.

Keywords: IM-MS, mass spectrometry, collision induced dynamics, drug metabolites identification, collisional cross sections, MOBCAL, Lennard-Jones potential, pharmaceutical industry.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

bash	Bourne Again SHell
CCS	Collisional Cross Section(s)
CDS	Collision Dynamics Simulations
CID	Collision Induced Dissociation
DT	Drift Tube
EHS	Exact Hard Spheres
ESI	Electro Spray Ionisation
GSK	GlaxoSmithKline
IL	Iteration Limit
IMS-MS / IM-MS	Ion Mobility Spectrometry - Mass Spectrometry
LJ	Lennard-Jones
MAD	Median Absolute Deviation
MCP	Multi-Channel Plate
MM	Molecular Modelling
MOE	Molecular Operating Environment
MS	Mass Spectrometry
NMR	Nuclear Magnetic Resonance
PA	Projection Approximation
QM	Quantum Mechanical
RF	Radio Frequency
RL	Rejection Limit
SMILES	Simplified Molecular-Input Line-Entry System

SVL	Scientific Vector Language
TM	Trajectory Method
ToF	Time-of-Flight
TW	Travelling Wave
UPLC	Ultra Performance Liquid Chromatography

Motivation

In the pharmaceutical industry, it is important to understand the absorption, distribution, metabolism and excretion properties of a drug in order to assess its pharmacology and safety. Drug metabolism is a major determinant for the changes in physiological drug concentration and can determine or alter its pharmacological or toxicological pathway [1]. It is therefore imperative to ascertain what happens to a drug in a living organism with a view of relating this to the action of the drug or its metabolites [2]. Drug metabolites are typically identified using a combination of techniques, but primarily the identification process starts with Ultra Performance Liquid Chromatography (UPLC). It is used to separate drug metabolites from endogenous species present in biological matrices. This permits more accurate Collisional Cross Sections (CCS) determination. After this, one of the following three techniques is used for further identification: High Resolution Mass Spectrometry (MS), Tandem MS and Nuclear Magnetic Resonance (NMR) [3, 4, 5]. Although MS can provide many structural clues to metabolites identity, it is generally not definitive, and only determines a mass of a sample or its parts but cannot distinguish between different isomers. This can be problematic when the exact structure of the metabolite is required to make an assessment of metabolite pharmacological activity or potential reactivity. In such cases robust metabolite isolation and subsequent NMR, both of which are time-consuming, are typically employed to generate unambiguous structural information. In addition to this, NMR is expensive and requires a lot of sample material [6, 7].

Collaborations between GlaxoSmithKline (GSK) and Waters MS Technologies Group have demonstrated the ability to differentiate between isomers of a drug based on their CCS using Ion Mobility Spectrometry - Mass Spectrometry (IMS-MS) and *in silico* modelling [8]. IMS-MS is a technique that allows separation of isomeric species based on differences in their CCS in the gas phase, thus providing specific information on the potential structure of a compound [9]. In combination with Molecular Modelling (MM), see Figure 0.1, it is considered as a potential tool for small molecule identification by measuring their gas-phase CCS and comparing them to theoretical CCS, derived using *in silico* approaches [8, 10]. Compared to NMR, IMS-MS requires less sample volumes, ultimately leading to reduced animal numbers in pre-clinical studies and the analysis of samples from lower dosed clinical trials. However, it requires high quality *in silico* methods to predict virtual CCS needed to elucidate the ones obtained experimentally.

A protocol for theoretical determination of CCS has been previously developed and introduced [11]. Consequently, a main focus of the project was refinement of the current workflow

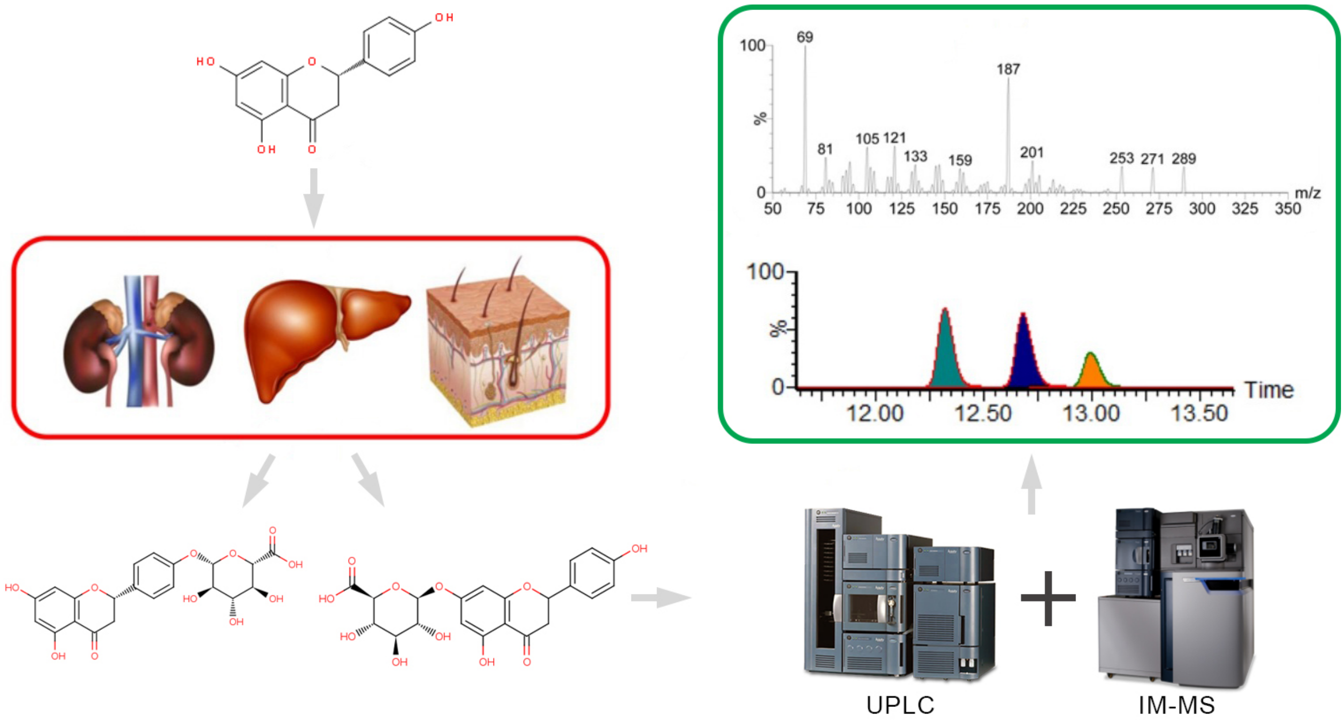


Figure 0.1 – A new approach towards drug metabolite structural isomer identification utilising IM-MS spectrometry. The method requires less time and sample material, while providing means to obtain unambiguous structural information. Experiments provide compound's m/z ratio and drift time to be later used in an empirical formula for CCS derivation.

and search of alternative methods for the improvement of a CCS calculation routine with the aim of reducing an error associated with its *in silico* prediction. Additionally, an existing version of the protocol suffered from being extensively time demanding, requiring human interaction at every step of its execution. Therefore, an automated alternative was desirable in order to provide an efficient means for large-scale modelling.

Having planned possible ways of improving the theoretical CCS calculation workflow, a closer look at other data available from experiments, particularly that of MS spectra from Collision-Induced Dynamics (CID), was of interest. Collision Dynamics Simulations (CDS) are capable of modelling CID processes by calculating an ensemble of trajectories, for which an ion of interest is colliding with buffer gas with given relative translational energy. Such simulations could potentially help to gain useful insights into the collision process by analysing obtained reactive trajectories, occurring within simulated fragmentation pathways [12].

1. Introduction

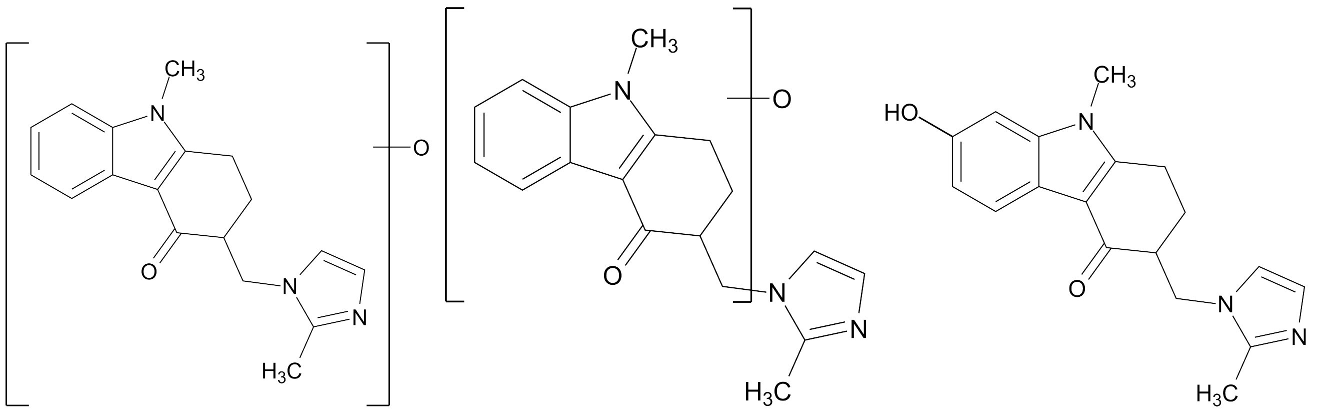
The project links various theoretical and experimental techniques, namely, IMS-MS, MM and, at a latter stage, Collision Induced Dissociation (CID). They formed a basis around which the project had been developing throughout its course along with additional resources (statistical software, open MS databases, etc.). While experimental data was mainly provided by a GSK Global Spectroscopy Department, theoretical values were obtained by running calculations on site.

1.1 Drug Metabolites Identification Techniques

Three approaches are normally available for drug metabolites identification, performed after UPLC: High Resolution MS (Figure 1.1a), Tandem MS (Figure 1.1b) and NMR (Figure 1.1c). Although MS can provide many structural clues to metabolites identity, it is generally not definitive, and only determines a mass of a sample but cannot distinguish among different isomers. In the case of Tandem MS, even though the technique allows to split a metabolite into parts and allocate the protonated part (as can be seen on Figure 1.1b), it is still not able to determine structural isomers. This can be problematic when the exact structure of the metabolite is required to make an assessment of metabolite pharmacological activity or potential reactivity. In such cases robust metabolite isolation and subsequent NMR, both of which are time-consuming, are typically employed to generate unambiguous structural information. In addition to this, NMR is expensive and requires a lot of sample material. Therefore, an alternative approach is desirable.

1.2 Ion-Mobility Spectrometry - Mass Spectrometry (IMS-MS)

IMS-MS is an analytical technique that offers potential for small molecule structural isomer identification by measuring their gas-phase collisional cross sections. It is being used for various purposes from studying properties of a particle in macromolecules, biomolecules, polymers,



(a) HRMS (b) Tandem MS (c) NMR

Figure 1.1 – Main drug metabolites identification techniques available on market.

etc. to detection of chemical warfare agents [13, 14], relying on the instrument's ion transmission ability and its separation capacity. Successful application of this approach could go beyond MS and can represent, in some cases, an alternative to NMR and reduce time and amount of sample required and so, reducing the animal use in research. Further development of this technique may provide an additional tool for structure elucidation to complement and/or supplement existing approaches. In favourable instances, this tool may enable the characterisation of drug metabolites without the need for metabolite synthesis or NMR structural assignment at all, or, at the very least, provide supportive data. It may benefit other areas of pharmaceutical structural characterisation as well.

1.2.1 Working principle

Integrating MS with IMS provides an extra dimension to unambiguous sample identification, yielding a three-dimensional spectrum (mass-to-charge ratio, intensity and drift time; see Figure 1.2). Structural features of an investigated ion are determined by measuring its arrival time distribution and mass-to-charge ratio after travelling through a background buffer gas (typically He or N_2) under the influence of a weak electric field. This separation technique allows to reduce a spectral overlap providing resolution of heterogeneous complexes with very similar masses, or mass-to-charge ratios, but different drift times. The latter ones provide an important layer of structural information — a CCS value, that can be calculated using a calibration curve generated from calibrant proteins with defined cross sections and is related to an overall shape and topology of the ion.

The identification process utilises an exponential correlation of the form, [16]:

$$\Omega \sim t_D^X, \tag{1.2.1}$$

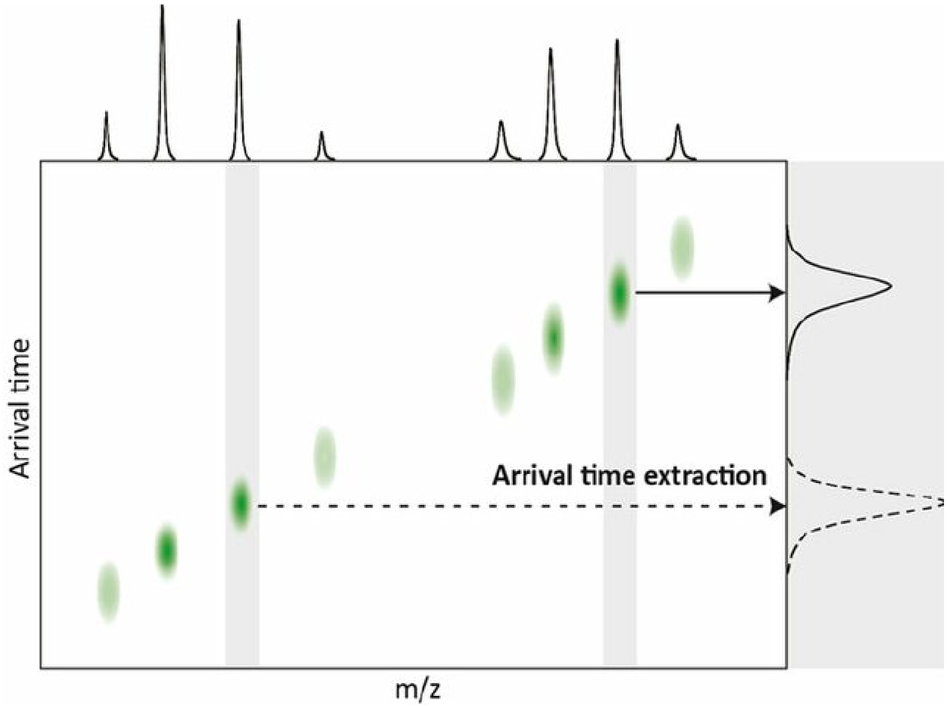


Figure 1.2 – Schematic IM-MS output spectra (via [15]). Results, produced by a combination of IM with MS, form a three-dimensional set of data, in which every feature, observed in m/z, has an associated arrival time distribution. Using this data, a CCS of a compound may be determined following steps described in this section. This information is particularly helpful for heterogeneous complexes, which populate multiple forms at equilibrium, as can be seen on the figure, and for which standard approaches would typically provide an ensemble average.

where t_D is measured experimental drift time and X is a constant obtained from a calibration curve. The procedure of producing the latter one is performed in the following sequence and is thoroughly described in [17] (for a Travelling Wave (TW) regime; more on different regimes in Subsection 1.2.4):

1. Calibrate drift time measurements by using a test set of compounds with known values of CCS. To do this, IMS-MS data is acquired for each test structure under exactly the same instrument conditions, that will be used for a target compound. Pressure and voltages must be kept identical to save the IMS separation settings. This provides experimental drift time values t_D for the test set.

2. Adjust each of the calibrant's experimental drift time t_D using an empirical formula:

$$t'_D = t_D - \frac{c\sqrt{\frac{m}{z}}}{1000}, \tag{1.2.2}$$

   where m/z is a mass-to-charge ratio for a given ion and c is an Enhanced Duty Cycle (EDC) delay coefficient [16], which is instrument-dependent.

3. Find corresponding CCS values for the test set in literature and/or in open databases and

apply a correction:

$$\Omega_C = \frac{\Omega}{z\sqrt{\dfrac{1}{m} + \dfrac{1}{M_G}}}, \tag{1.2.3}$$

where Ω_C is the corrected CCS, m is a molecular weight of an ion and M_G is a molecular weight of the buffer gas.

4. Make a plot $\ln t_D$ against $\ln \Omega_C$, which can be approximated as an equation of the form:

$$\ln \Omega_C = X \ln t'_D + A, \tag{1.2.4}$$

where X and A are obtained by fitting the plot to a linear relationship. X corresponds to the exponential constant from Equation (1.2.1), while A will be used at a later stage.

5. Check if a fit correlation coefficient r^2 is greater than 0.95:

$$r = \frac{\sum(x-\overline{x})(y-\overline{y})}{\sqrt{\sum(x-\overline{x})^2(y-\overline{y})^2}} \tag{1.2.5}$$

If the obtained value is not satisfying — repeat the measurements changing experimental samples and/or conditions until an appropriate value of r^2 is produced.

6. Readjust the calibrant drift time by making use of a newly calculated coefficient X:

$$t''_D = z t'^X_D \sqrt{\frac{1}{m} + \frac{1}{M_G}} \tag{1.2.6}$$

As an additional validation step, plot Ω_C vs t''_D and recalculate the fit correlation coefficient using Equation (1.2.5). Similarly to Step 5, if $r^2 < 0.95$, one must rerun the measuring part.

7. Next, run measurements on target compounds. Correct measured t_D values using Equation (1.2.2), as it was done for the test set in Step 2.

8. Using Equation (1.2.6) and the value of the constant X, calculate the final value of the drift time t''_D for each of the target ions.

9. Finally, an experimental CCS value may be calculated as follows (A was obtained in Step 4):

$$\Omega_{exp} = A t''_D \tag{1.2.7}$$

Measurements for each target complex are run at least three times, after which a final value of CCS is calculated. This value is further compared to a corresponding theoretical result.

1.2.2 Experimental Setup

Two experimental setups are commercially available and are widely used for ion identification studies with IMS-MS: a Synapt G2 HDMS [18] and a modified Synapt G1 HDMS (both designed by Waters, Manchester, the UK). The Synapt G2 HDMS, shown on Figure 1.3, is an IMS-MS instrument equipped with an Electro Spray Ionization (ESI) source, marked as INTELLISTART. Ionisation is an important part for studying structures with IMS-MS tools. There is a broad range of methods available on the market to ionise particles: Electron Ionisation, Thermospray, Chemical Ionisation, Atmospheric Pressure Ionisation, ESI, etc. In experiments of interest, ESI has been used for ionisation of chemical structures. The technique, developed by Fenn *et al* [19, 20], is utilised to prepare multiply charged structures of a given solution and is based on applying a strong electric field to liquid (a structure of interest + a buffer) under atmospheric pressure.

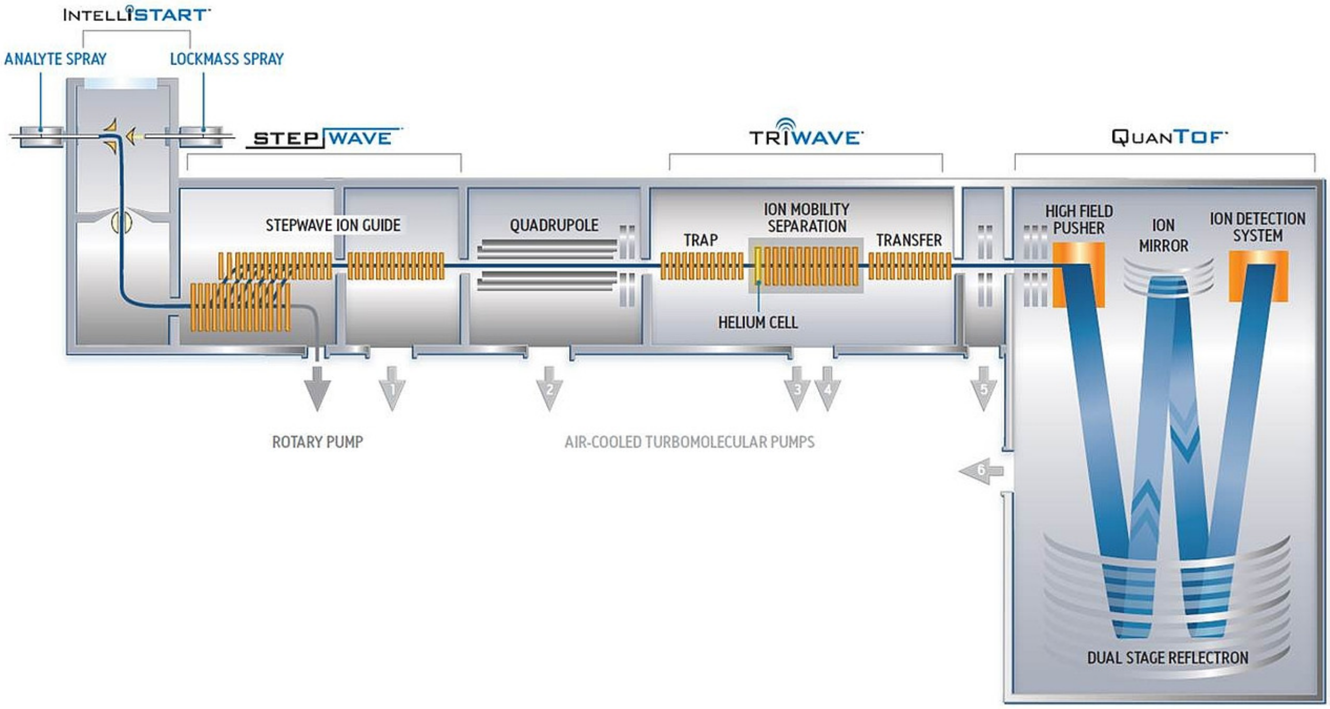


Figure 1.3 – Waters SYNAPT G2-Si High Definition Mass Spectrometer scheme (via [18]).

The Synapt G2 HDMS tool may be run in one of the two regimes: an IM-MS mode and an MS only mode (more details about the regimes in Subsection 1.2.4). After the ESI step, parent ions can be identified by a quadrupole mass filter (between STEPWAVE and TRIWAVE parts; explained in Subsection 1.2.3). If an IMS-MS regime is used, ions are gathered in a trapping area TRAP, filled with Ar (right after the QUADRUPOLE part of the setup); it is located directly in front of the ion-mobility analyser, denoted as ION MOBILITY SEPARATION. The first part of it contains He (HELIUM CELL) to ease transfer of the ions from the trap to the Ion-Mobility Separator, whose major part is filled with N_2. Here, ions with the same mass are separated according to their varying drift times. After that, the ions pass through the TRANSFER section, where a Time-of-Flight (ToF) mass spectrometer (QUANTOF) performs analysis. It has a mass resolution up to 40000 (50% valley definition). As a final step, collected data is transferred to a computer, where an $Intensity$ vs m/z spectrum is analysed.

Additionally, this facility allows ion-molecule reactions studies by using the trap, IM and/or the transfer cell as reaction cells, respectively. Individual modifications to the setup are also possible.

1.2.3 Quadrupole and Time-of-Flight (ToF) analyser

A quadrupole analyser deploys an oscillating electric field to distinguish and separate ions, based solely on their m/z ratio, by judging their trajectories stability in the present electric field. It consists of four perfectly parallel cylindrical rods (two sets of symmetrically opposite rods of the same charge, as can be seen on Figure 1.4), aligned in a way that allows some space between the rods for ions to travel through freely. There are two types of voltage supplied. One is Direct Current voltage, and other is a superimposed RF voltage (causes the ions to spiral as they traverse the quadrupole towards a detector). Such arrangement creates a continuously varying electric field along the length of the analyser. If a particle is negatively charged, it will try to move towards the positive electrode and, before it gets discharged to the electrode, a polarity is changed. This varying electric field is precisely controlled so that during each stage of a scan, ions of a particular mass-to-charge ratio can only pass down the length of the analyser (due to their stable trajectories), whereas others are therefore eliminated.

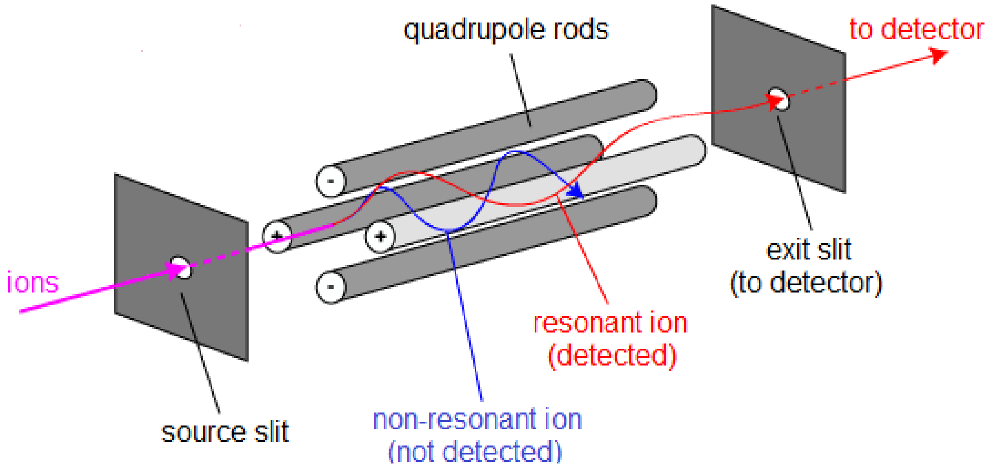


Figure 1.4 – A typical quadrupole setup. After passing through the source slit, ions travel in the space inside the four rods. Depending on a chosen m/z ratio, some part of them will successfully traverse the region and will be detected (resonant ions), while the rest will be neutralised by one of the rods (non-resonant ions).

Quadrupoles may be arranged in various ways, depending on a desired type of analysis and/or scan. They are generally used for better transmission efficiency. A typical experimental IM-MS setup may involve a few of them; for example, after the first such device, collision gas is normally introduced in the second quadrupole, whose function is to provide ion fragmentation. Additionally, a third quadrupole may be added to select specific ions, targeted for further analysis. Time-of-Flight (ToF) instruments are responsible for segregating ions based on their m/z ratio. They have fast, precise electronics and modern ionization techniques, like ESI, mentioned in Subsection 1.2.2.

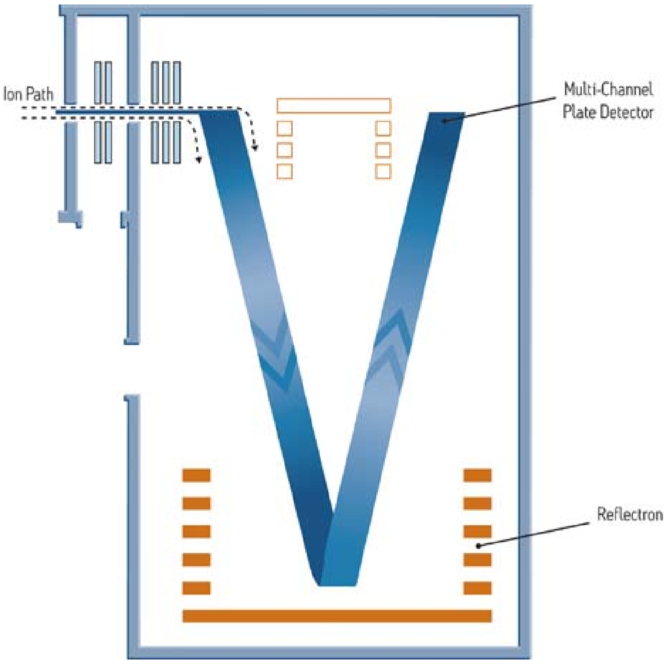


Figure 1.5 – In ToF, ions are injected via an *Ion Path* and are accelerated by a high voltage pulse into a flight tube (blue), where they strike a *Reflectron*. Lighter ions arrive at the *Multi-Channel Plate (MCP) Detector* sooner than heavy ones.

A ToF analysis starts by accelerating a group of ions, injected via an *Ion Path*, in an instant burst at the same voltage, towards a *Multi-Channel Plate (MCP) Detector* (see Figure 1.5, via [21]). The ions escape a source, each having received from a "pusher" electrode an identical electrical charge (potential). After that, they travel into a very low pressure tube. Since kinetic energy of similarly charged compounds will be the same ($E_k = {}^1\!/\!{}_2 mv^2$, where m is ion's mass and v is its velocity), those with lower mass to charge ratios will experience greater velocity and a shorter interval before striking the MCP detector. Finally, since E_k, m and z determine ion's arrival time to the detector, one can deduce that:

$$v = \frac{d}{t} = \sqrt{\frac{2E_k}{m}}, \tag{1.2.8}$$

where d is a distance a particular ion travels during time t, which in turn depends on the m/z ratio. And, by expressing t from Equation (1.2.8), one gets:

$$t = d\sqrt{\frac{m}{2E_k}} \tag{1.2.9}$$

The described principle provides an accurate measurement of masses at different time scales in a well-calibrated ToF setup. Since all masses are measured for each injection of ions in the instrument, the ToF tool can achieve a very high sensitivity relative to scanning instruments. Apart from that, the technique offers spectral continuity, fast acquisition rates and a wide dynamic range without sacrificing speed or sensitivity.

Some experimental facilities, like ion traps, offer a combination of these capabilities. But until the implementation of hybrid instruments, such as IMS-MS, involving quadrupoles and ToF devices, no single tool could deliver high-order performance in all aspects.

1.2.4 Drift Tube vs Travelling Wave Modifications

It is generally accepted that a default mode for IMS-MS experiments is a Travelling Wave (TW) mode. It is overwhelmingly true for industry and academia. A Drift Tube (DT) mode is a rather trivial approach for gas-phase CCS measurements since it is based on first principles, laid out in [22]. Within it, a constant homogeneous potential gradient is applied along the tube (Figure 1.6 (A)), whereas in the TW mode (Figure 1.6 (B)) the ions are confined by a Radio Frequency (RF), applied to a stacked ring ion guide [23].

Simultaneously, a direct current voltage wave is travelling to the exit (T-wave, and, thus, the name of the technique). Higher mobility ions (green colour) are pushed more easily by the waves, while larger ions, and, therefore, less mobile (orange colour), have more friction with the background gas and, as a result, slip more often behind the waves and take longer to traverse the mobility cell.

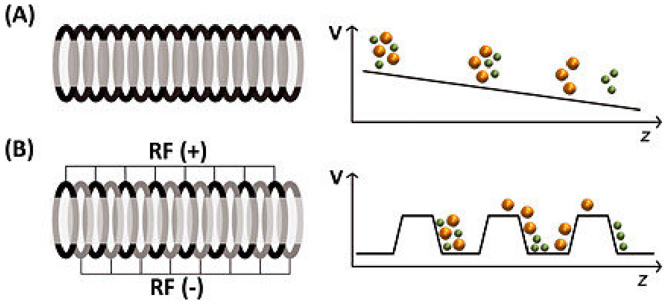


Figure 1.6 – Main difference between DT-IM-MS (A) and TW-IM-MS (B). High mobility ions are in green and low mobility ions are in orange.

Therefore, a T-Wave ion mobility device is more complex than a DT one since the electric field is not constant so direct determination of mobilities from the measured drift times is not straightforward. This requires, as it has been described in Subsection 1.2.1, the T-Wave mobility separation to be calibrated using species of known CCS determined using standard drift tubes to provide meaningful CCS results routinely. Therefore, the calibration process amounts to main sources of errors in the methodology [16]. Among other possible errors are: disagreement between gases, used in deriving Helium CCS values with the TW mode (run with Nitrogen gas), but with data used for calibration being obtained with the DT mode and Helium as a buffer gas; and systematic errors due to the influence of a travelling wave electric field regime [24, 25]. Typically, accuracy of experimental CCS data, provided by the TW modification varies in the range $2-5\%$ for proteins and small molecules [9, 26].

To conclude, measurement of ion drift time with IMS-MS provides, via an empirical relation, a CCS for a given compound. On the other hand, this value can be obtained by a computational protocol which has been previously developed [11]. Therefore, only a combination of the two approaches — the experimental and theoretical ones — allows to identify metabolites as IM-MS does not provide structural information, if used independently, but only a CCS value, whereas the computational protocol can help to deduce structural information knowing compound's CCS.

1.3 Collisional Cross Sections (CCS)

When an ion moves through a drift tube, it gains some average drift velocity v, which is defined as a product of an electrostatic field E and ion mobility K:

$$v = KE \tag{1.3.1}$$

The zero-field mobility K in a gas-filled cell is a measure of how rapidly the ion moves under the influence of some uniform electric field E, taking into account ion's repeated collisions with neutral buffer gas (illustrated on Figure 1.8). It depends on the charge state and the shape of the ion. Ions of higher charge states will experience a stronger drift force, while traversing the drift tube, leading to their shorter drift times (Figure 1.7[1], left).

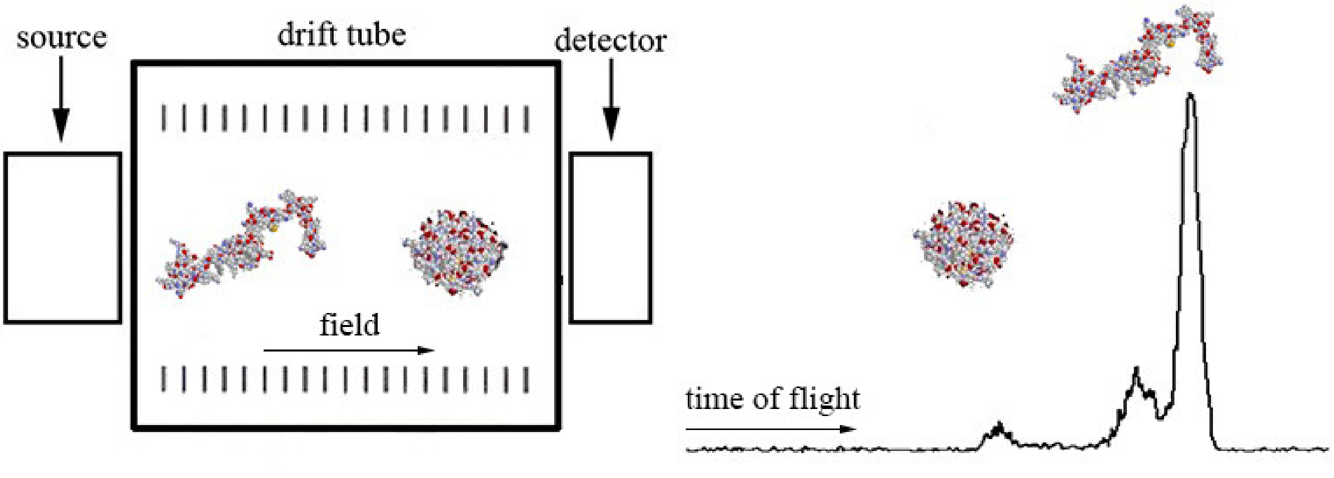


Figure 1.7 – Experimental mobility separation for an $[M + 8H]^{8+}$ ion of cytochrome *c*. As shown here, a higher charge state ion (a sphere-like structure) travels through a drift tube faster than singly charged compounds (left), meanwhile ion conformations, that are more elongated, have lower mobilities and longer drift times relative to those with a more compact conformation (right).

At the same time, compact ion conformations or ions, that have more spheric or plain structures, will have shorter drift times as they will experience a smaller number of collisions rather than the ones with any concave or convex features present (Figure 1.7[1], right).

[1] via `http://www.indiana.edu/~clemmer/Research/Intro.php`

The mobility K is defined as [27]:

$$K = \frac{\sqrt{18\pi}}{16}\sqrt{\left[\frac{1}{m} + \frac{1}{m_{gas}}\right]}\frac{ze}{\sqrt{k_B T}\Omega_{avg}^{(1,1)}}\frac{1}{N} \tag{1.3.2}$$

As one can observe from Equation (1.3.2), K depends on mass m, a charge z and shape of the ion, incorporated by an orientationally averaged collision integral $\Omega_{avg}^{(1,1)}(T)$, used further to obtain a CCS. $\Omega_{avg}^{(1,1)}(T)$ is an amount of momentum transferred to a nanoparticle by the impingement of surrounding gas molecules, leading particles of different size and structure to migrate differently through a background gas. It can be expressed as shown in Equation (1.3.3) and approximations and complex integrations are required for its theoretical calculation:

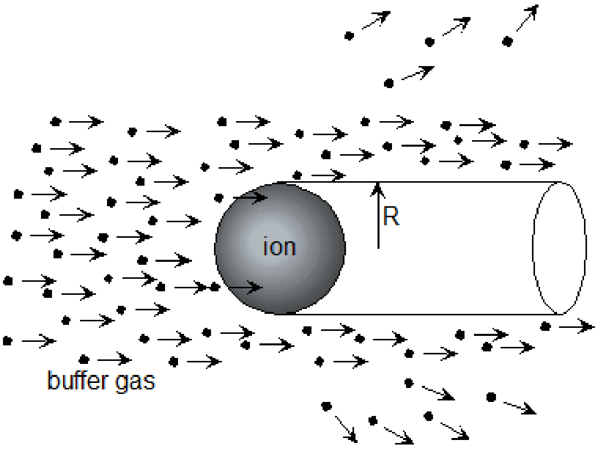


Figure 1.8 – A schematic representation of an ion interaction with buffer gas molecules. The particles, that have been scattered, would have been located in the cylinder, had the ion not been there. The larger a CCS value is — the more particles are scattered, leading to more interactions.

$$\Omega_{avg}^{(1,1)}(T) = k\int_0^\infty \mathrm{d}\epsilon\ f(\epsilon, T)\sigma(\epsilon,\chi)\int_0^\pi \mathrm{d}\chi\ (1-\cos\chi) =$$
$$\frac{1}{8\pi^2}\int_0^{2\pi}\mathrm{d}\theta\int_0^{\pi}\mathrm{d}\phi\ \sin\phi\int_0^{2\pi}\mathrm{d}\gamma\ \frac{\pi}{8}\left(\frac{\mu}{k_B T}\right)^3\int_0^\infty \mathrm{d}g\ g^5 e^{-\mu g^2/2k_B T}\int_0^\infty \mathrm{d}b\ 2b(1-\cos\chi(\theta,\phi,\gamma,g,b)), \tag{1.3.3}$$

where k is a normalisation constant, ϵ is kinetic energy of the buffer gas, T is temperature, μ is a reduced mass, g is a relative velocity and χ is a scattering angle (the angle between the ion's trajectory before and after a collision with a buffer gas atom takes place). The latter one defines the extent of momentum transfer during a collision, and, therefore is the most crucial determinant for accurate CCS calculation. Unfortunately, one cannot obtain scattering angles analytically, apart from some very rare simple cases, due to highly non-local and correlated nature of the van der Waals interaction between an analyte ion and the buffer gas. As a result, they need to be derived numerically using detailed information related to an ion-neutral interaction potential. θ, ϕ, γ are three angles that define the collision geometry. The term $f(\epsilon, T)$ denotes a distribution of kinetic energies (velocities) at the temperature T (i.e. Maxwell distribution), whereas $\sigma(\epsilon,\chi)$ is a probability distribution of the deflection angle for trajectories with kinetic energy ϵ. $\sigma(\epsilon,\chi)$ can be obtained by solving equations of motion for a sufficient large number of collisions. The key point is that thanks to different shapes of ions, isomers can be identified. Additionally, to be distinguished by IM-MS, isomers must conform to energetically distinct conformations. Finally, it is important to keep in mind, that originally CCS is linked to momentum.

In addition to this, two effects must be taken into account while developing a theory for collision integral calculation: *size* and *shape* effects [28]. The *size* effect occurs in analyte ions of a medium size (tens to thousands atoms), for which glancing collisions become increasingly important (with increasing molecular size) as a consequence of a deepened interaction potential. In other words, an effective atomic radius of a particular atom in a compound increases due to interaction with its surrounding molecular environment. Moreover, analyte's molecular geometry (curvature) also influences on the interaction potential; this is known as the *shape* effect. It is the most evident for concave features of compounds, e.g. cups, where the deflection angle can become very large (up to 180° for a large number of collision geometries) and can result in high momentum transfer. The fundamental cause for both effects lies in the simultaneous interaction of buffer gas electrons with electrons of the compound. These effects are non-local (collective) and are coupled to each other.

1.3.1 Assumptions

Experimental evidence suggests that in the majority of measurements with the most commonly used background gases, such as He, air and N_2, gas molecules re-emission is largely inelastic in its nature (with exchange of energy among rotational, vibrational and translational modes). However, if one wants to calculate CCS computationally, this level of complexity of the involved processes has to be reduced by introducing a few assumptions:

- no changes in vibrational or rotational energies of an ion occur;
- collisions between gas molecules and compounds are completely elastic (translational energy of gas molecules is conserved);
- reflections are regarded as specular (the angle of reflection is the same as that of the angle of incident);
- all compounds are treated as rigid bodies. Therefore all conformations should be taken into account and an average value of them has to be found.

Various theoretical approaches for collision integral $\Omega_{avg}^{(1,1)}(T)$ calculation, mimicking ion-neutral collision processes and, thus, leading to a description of the ion-buffer gas interaction, have been developed. The most common ones include a Projection Approximation (PA, [29]), an Exact Hard Spheres (EHS, [30]) and a Trajectory Method (TM, [31]). These techniques make use of different approaches, such as neglecting *size* (EHS as in [31] and [32]; PA as in [28]) and/or *shape* (PA as in [28]) effects or introducing an interaction potential to describe the system (TM as in [31]). The accuracy of these models is largely dependent on the empirical parameters used for ion-buffer gas interactions. A brief theoretical overview is given in the following Sections 1.3.2–4.

1.3.2 Projection Approximation (PA)

Projection Approximation is the default method for the protocol under study (Chapter 2). Its development dates back to 1925 by Mack [29] and introduces a fairly simple concept. If one takes a lantern and shines with it at an object against a plane, one will see the object's projection on that plane. Doing so from as many directions as possible will result into a set of projections and, for each of them, a projected area can be calculated. After that an average can be found. This was the initial idea introduced in the paper [29], where the author used a strong beam of parallel light, coming from a stereopticon lantern, along with a lens system, to elucidate beeswax, mounted on a special device (Figure 1.9, via [29]). The setup allowed to position the beeswax at any desired angle with respect to the light beam, directing its shadows on a paper screen. The outline of each shadow was traced with a pencil and its area computed.

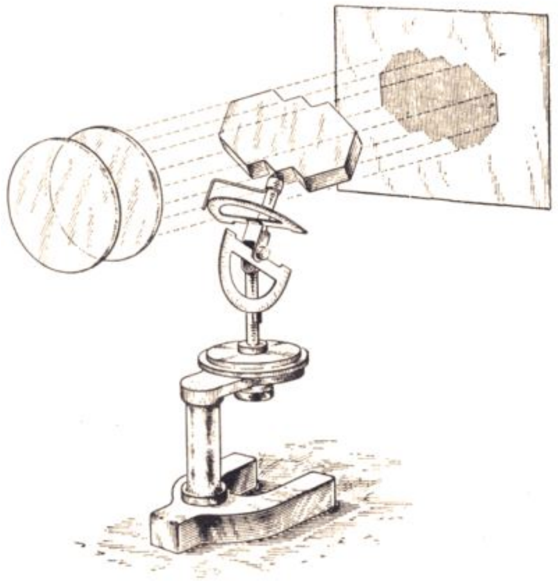


Figure 1.9 – An original sketch of a device used by Mack to fix a beeswax between a light source and a paper wall, where shadows outlines were registered.

Since every orientation of the beeswax was equally probable, an average shadow area (which is an average cross section) was calculated as a total area of the shadows divided by their total number. The same approach can be applied if one has a compound: obtaining its projections on planes would allow to calculate its CCS.

The PA method considers polyatomic ions as collections of spheres of radius R_{coll}, as on Figure 1.10. It represents only three possible orientations, however, in a case of non-spheric surfaces (which is the most common scenario), a larger number of orientations can be produced to obtain a more precise value of a CCS. Consequently, a collision integral $\Omega_{avg}^{(1,1)}$ is defined as an orientation averaged area of a whole set of these spheres, forming the ion, projected on a plane, perpendicular to orientation axes:

$$\Omega_{avg}^{(1,1)} = \frac{1}{n}\sum_{i=1}^{n}\Omega_i, \qquad (1.3.4)$$

where n is a total number of projections collected and Ω_i is a value of a particular projection i.

This model is entirely local and completely neglects both *size* and *shape* effects. Therefore any convex and/or concave features of compounds are lost. Additionally, PA is not capable to include long-range interactions. Moreover, it can be deduced that within this method's approximations, any details of scattering processes do not play any role in evaluation of $\Omega_{avg}^{(1,1)}$ and, as a result, in CCS determination, as PA ignores them. This makes it computationally highly efficient, but may lead sometimes to inaccurate results due to introduced simplifications [28].

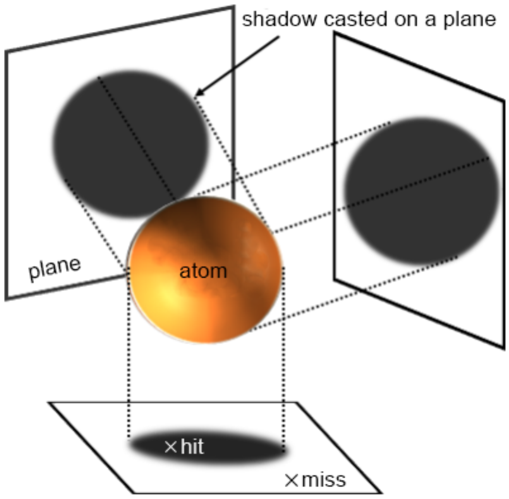


Figure 1.10 – Within the PA method, ions are represented as a collection of spheres (atoms), and their CCS is found as an averaged projection area.

The most popular computational method for a PA model implementation is based on Monte Carlo simulations. They are performed by first defining an effective hard sphere radii for all atoms in a structure of interest and for buffer gas molecules, considered for a collision. This creates a region or a domain, which encompasses the whole structure inside. Then the collision process is simulated by randomly modelling hits on a plane with the domain's projected area (bottom plane on Figure 1.10), that is, if a centre of mass of a gas molecule with a radii r_{gas} falls inside the projected area, it is a hit, otherwise it is registered as a miss. Next, an approximate area of the structure's projection (its CCS) is given as:

$$\Omega_{PA,i} = \frac{N_{hits}}{N_{total}} A_{plane} = \frac{N_{hits}}{N_{hits} + N_{misses}} A_{plane}, \tag{1.3.5}$$

where N_{hits} is a number of hits inside the projected area, N_{misses} is a number of misses, N_{total} is a total number of hits and A_{plane} is the area of the plane. After that the whole molecule (the domain) is rotated around its centre of mass using uniform random rotation matrices, constructed from Euler angles [33] (Figure 1.11 demonstrates a few examples of different orientations of the same molecule and their corresponding projections, via [34]). It is necessary to explore ion's rotational space. The next step is to generate a new domain's projection, followed by a subsequent "hit-and-miss" simulation. Once a sufficient number of $\Omega_{PA,i}$ values is collected, a final value of Ω_{PA} is found by computing their average. There are other similar methods available, like a Projection Superposition Approximation [28], which is more refined than PA, although it did not find a widespread use.

1.3.3 Exact Hard Spheres (EHS)

A higher level of complexity is introduced by an Exact Hard Spheres (EHS) method. As the name suggests, EHS considers molecules as a collection of spheres with some radii r_i, in the similar fashion as the PA method does, but additionally employs an infinite hard wall potential between colliding particles taking into account a scattering effect.

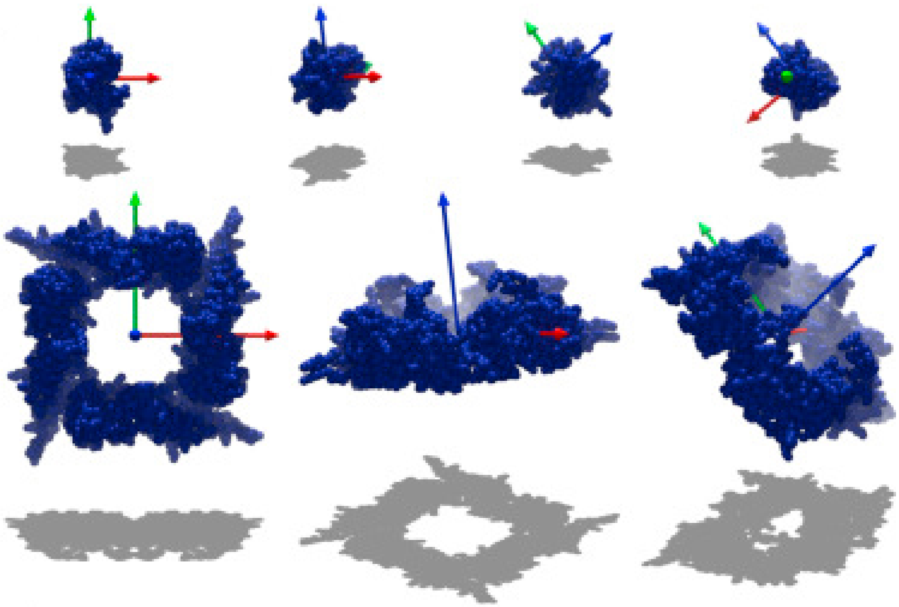


Figure 1.11 – The PA method calculates CCS as a rotational average of target's projected area, adjusted for the finite radii of gas molecules. The target is rotated randomly around its centre of mass many times to explore rotational space and the average projected area is determined through Monte Carlo sampling.

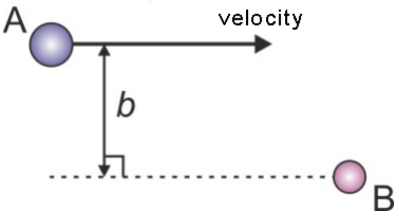


Figure 1.12 – Explanation of an impact parameter b in the EHS model.

The hard wall potential means that impinging gas molecules will re-emit as soon as a collision between them and the structure of interest takes place whilst they cannot penetrate the structure. From Figure 1.13, left, one can deduce that a collision only occurs when the particles meet at a contact distance b_{min} (which is a sum of the radii of colliding hard sphere particles in the EHS model). This is a so called impact parameter b, which describes an initial perpendicular separation of trajectories of the collision partners (Figure 1.12). Essentially, this is a distance at which the colliding pair would miss each other if they did not interact at all, and can be found by extrapolating the initial straight-line paths of the particles at large separations to the distance of closest approach. b_{min} simply determines a separation distance at which a collision is unavoidable. Thus, potential energy of a colliding pair depends solely

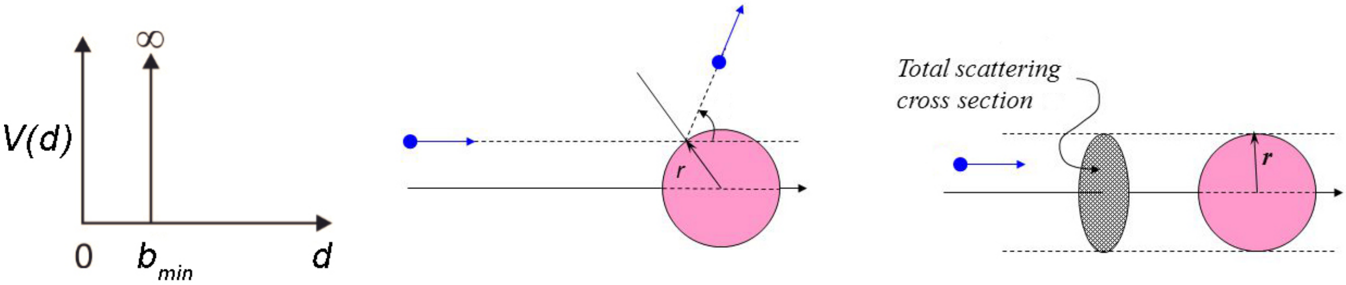


Figure 1.13 – Illustration of an infinite hard wall potential (left) and of a collision process within the EHS model (centre and right).

on a single coordinate, namely on their effective impact parameter. As mentioned in Subsection 1.3.1, such collisions are specular and elastic (Figure 1.13, centre). In the description of their model in [30], Shvartsburg and Jarrod define an averaged collision integral for an arbitrary body as (can be derived from Equation (1.3.3)):

$$\Omega_{avg}^{(1,1)} = \frac{1}{4\pi^2}\int_0^{2\pi} \mathrm{d}\theta \int_0^{\pi} \mathrm{d}\phi \sin\phi \int_0^{2\pi} \mathrm{d}\gamma \int_0^{\infty} \mathrm{d}b 2b(1-\cos\chi(\theta,\phi,\gamma,b)), \tag{1.3.6}$$

where θ, ϕ, γ describe the collision geometry and b is an impact parameter. This equation can be solved numerically if radii of involved atoms are known. To define a scattering angle χ, a simulated trajectory is followed through all collisions with its environment until the trajectory leaves it. This allows to account for multiple collisions from one part of the cluster, where the body is, to another part. In a case of an ideally sphere-like molecule, for example, simulated as a hard sphere, a CCS would be equal to a molecule's projection, have a shape of a circle (Figure 1.13, right) and would be defined as following, via [30]:

$$\Omega_{avg}^{(1,1)} \approx \frac{1}{4\pi^2}\int_0^{2\pi} \mathrm{d}\theta \int_0^{\pi} \mathrm{d}\phi \ \sin\phi \int_0^{2\pi} \mathrm{d}\gamma \ \pi b_{min}^2 = \pi b_{min}^2 \tag{1.3.7}$$

EHS treats collisions independently of the number of present spheres or atoms, representing the molecule (molecule size), and independently of the arrangement of the spheres (molecule geometry). Therefore, the size effect is not included, just like in the PA model. On the other hand, it does approximate the non-local shape effect by allowing multiple collisions to occur within the course of one trajectory.

Similarly to the PA model, despite including scattering effects and collision processes, EHS does not introduce the effects of long range potentials between the background gas and the molecular ion. Eventually, the model did not enjoy high popularity and was largely overshadowed by faster or more precise alternatives (the PA and TM models, respectively).

1.3.4 Trajectory Method (TM)

A Trajectory Method (TM) is another way of calculating CCS, taking a more rigorous approach than the methods, described in the previous subsections. It was first introduced by Mesleh *et al* [31] in 1996 and gained a widespread recognition as the most accurate method for CCS calculations. It is based on the claim that short- and long-distance interactions can be described using a Lennard-Jones potential (an LJ potential).

The potential, developed in the original work, had a following form [31]:

$$\Phi(\theta,\phi,\gamma,b,r) = 4\epsilon\sum_{i}^{n}\left[\left(\frac{\sigma}{r_i}\right)^{12} - \left(\frac{\sigma}{r_i}\right)^{6}\right] - \frac{\alpha}{2}\left(\frac{ze}{n}\right)^{2}\left[\left(\sum_{i}^{n}\frac{x_i}{r_i^3}\right)^{2} + \left(\sum_{i}^{n}\frac{y_i}{r_i^3}\right)^{2} + \left(\sum_{i}^{n}\frac{z_i}{r_i^3}\right)^{2}\right], \tag{1.3.8}$$

where ϵ and σ are Lennard-Jones potential parameters: ϵ is the well's depth and σ is a distance at which the potential becomes positive).

α denotes the polarizability of the buffer gas, ze is a charge of a particular atom in the compound, n is a total number of the atoms and, finally, r_i, x_i, y_i, z_i are coordinates that define spacial orientation of the compound's atoms with respect to the buffer gas atoms. The first term in Equation (1.3.8) is a 12-6 repulsive-attractive Lennard-Jones potential, seen on Figure 1.14, and the second term is ion-induced dipole interaction, that takes into account emergence of dipoles due to the presence of charged ions. A contribution of the latter term decreases with increasing system size. Within the method, a Runge-Kutta-Gill initiator and an Adams-Moulton predictor-corrector propagator were deployed to calculate trajectories with energy conservation for all trajectories better than 0.5% [31]. A schematic visualisation of trajectories computed with the TM method are shown on the Figure 1.15: the trajectories in the closest proximity to the ion will collide with it, whereas the ones further away will have their travelling paths affected due to the interaction potential.

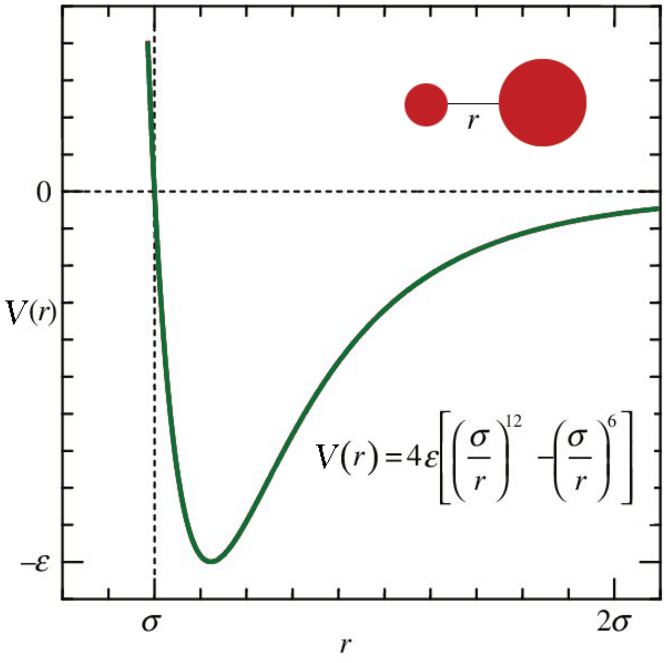


Figure 1.14 – A standard Lennard-Jones interaction potential — the most used realistic model of a two particle interaction. The first term describes a short-range repulsive interaction, whereas the second — long-range attractive interaction.

Due to numerous force evaluations for every simulated trajectory, embedded with large integration domains, run for many systems, TM becomes very computationally expensive and extremely time consuming, losing in efficiency to the PA and EHSS methods, but providing a more reliable output. Surprisingly, the PA method occasionally yields results within a few percents of TM values, as reported by several papers ([28, 34, 35] and others), either due to the effects of a calibration procedure or thanks to accidental cancelling out of the effects of introduced approximations.

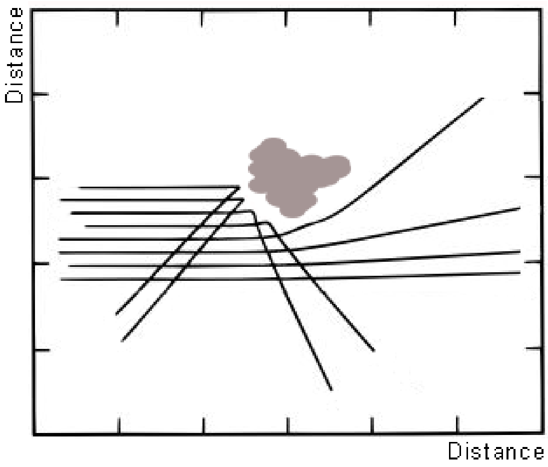


Figure 1.15 – A schematic visualisation of trajectories computed with the TM method. The trajectories in the closest proximity to the ion will interact with it, while the rest will be diverted due to the interaction potential.

Even though TM calculations are based on the most rigorous treatment, their accuracy heavily relies on the empirically determined LJ parameters ϵ and σ, defined in software used for CCS calculations, and, as it might happen, an empirically optimized method may show limited success for compounds not included in the initial test set. Since there is no one well-defined set of the mentioned quantities, applicable to any system, an additional refinement is sometimes required to obtain acceptable results. Several successful attempts have been made to optimise TM parameters by Wu, T. *et al* [36] and Campuzano *et al*[9], showing that indeed an improvement can be achieved by adjusting implemented in the TM method parameters. Thus, two methods for theoretical CCS calculations have gained broad popularity — the PA and TM methods, and often both of them are used, since the PA method is very quick and including it gives an additional set of results in conjunction with TM, allowing to make a comparison.

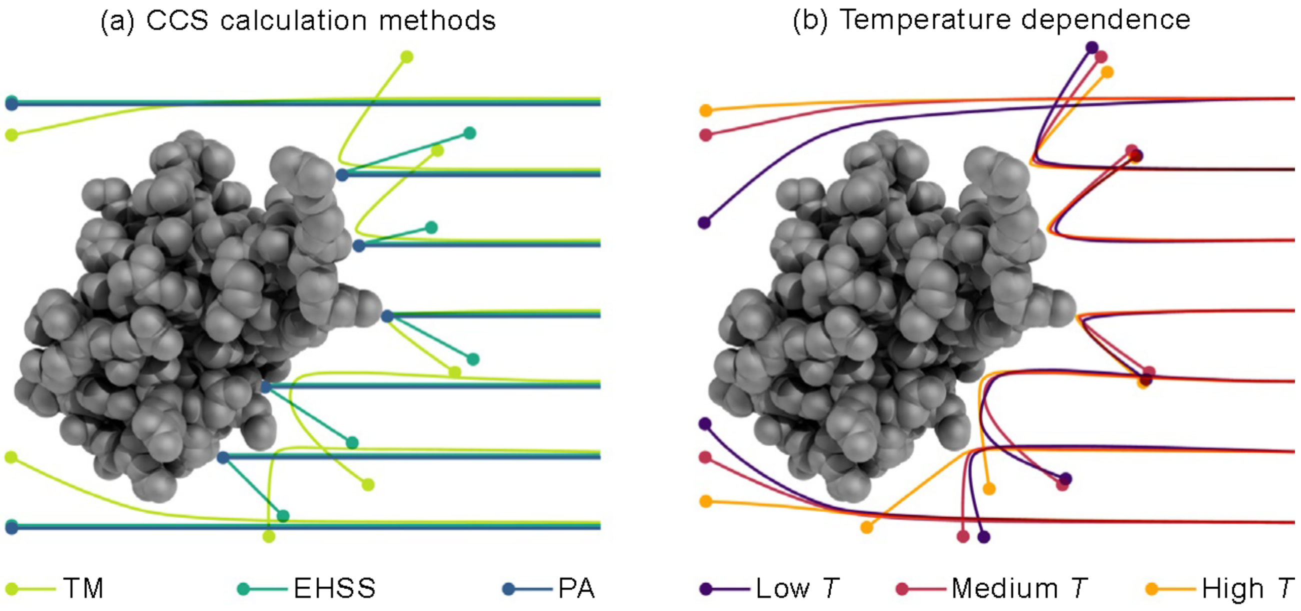


Figure 1.16 – Analogy between theoretical methods for CCS calculation and experiments, run under various temperatures. Both illustrations help to understand underlying approximations behind the introduced models. (a) depicts how collisions happen under the PA, EHSS and TM models, whereas (b) resembles temperature dependence of the collision process.

Figure 1.16 summarises the information discussed in Sections 1.3.2–4 and gives a visual explanation on how different CCS calculation methods work and demonstrates the effect of temperature on collision process during an experiment (via [37]). Figure 1.16 (a) shows how a collision happens under the PA, EHSS and TM models: PA accounts for direct contact between background gas and a compound, but only distinguishes hits and misses; EHSS, similarly to PA, works with direct interactions, however, additionally considers multiple scattering after the initial collision. Finally, TM thoroughly follows the background gas particles' trajectories. On the other hand, experimental collision process is largely dependent on temperatures, as shown on Figure 1.16 (b): higher temperatures lead to decreased momentum transfer between buffer gas and the compound due to higher velocities of the buffer gas particles and, thus, shorter interaction times.

1.3.5 MOBCAL Software

All described above theoretical CCS calculation methods have been employed in the MOBCAL (MOBility CALculation) software[2]. It was developed by Jarrold M. F. *et al* [30, 31] at University of Indiana and is a broadly used programme for CCS calculations in the field. It was written in the FORTRAN77 programming language and various versions of it have been independently developed by scientists. Other packages are now available on the market with similar or more advanced and efficient functional codes, including parallelised codes. For example, IMoS[3] [38] or Collidoscope [39] (directly based on MOBCAL) are available to researchers, however, MOBCAL has become a first choice in the project throughout its development and is still being used. Nevertheless, the use of such advanced tools is encouraged and may be possibly explored in further studies. The original source code of MOBCAL requires *.mfj* files and does not accept a *.pdb* format, however, the software has been modified by Dr. Jordi Munoz-Muriedas to accept structures, contained in *.pdb* files, along with a few others crucial routines (used in [11]). In addition to this, further modifications of MOBCAL have been made throughout the development of the project.

Remark 1.3.1. *Technical support for MOBCAL is no longer provided, therefore all changes to the source code were made purely for the purpose of this project.*

1.4 Collision Dynamics Simulations (CDS)

Studies of many biomolecules are often performed in gas-phase, since such conditions eliminate environment-dependent effects and provide direct access to crucial intrinsic properties. In combination with ESI (briefly mentioned in Section 1.2.2), Tandem MS with the help of Collision Induced Dissociation of ions (CID) builds up a reliable approach, able to deliver understanding of reactivity and characteristic features of biomolecules [40]. This is due to extensive fragmentation

[2] https://www.indiana.edu/~nano/software/

[3] http://www.imospedia.com/imos/

that occurs in activated ions, provided they have sufficient amount of energy. A schematic representation of what happens during an MS/MS measurement involving CID, is shown on Figure 1.17: an ion source injects a collection of compounds among which a mass spectrometer selects only the compound of interest. The latter one undergoes CID and its fragments are passed to a second mass spectrometer, which identifies the fragments and sends them to a detector. Based on the information received, an experimental MS spectrum is built. There are other techniques available for such ion activation, varying in instrumentation, underlying mechanisms and energy range used, however, the combination of ESI, Tandem MS and CID has proved to be one of the most reliable [41, 42, 43]. Therefore, its theoretical counterpart was included in this PhD project with an aim of getting more insights into studied collision processes.

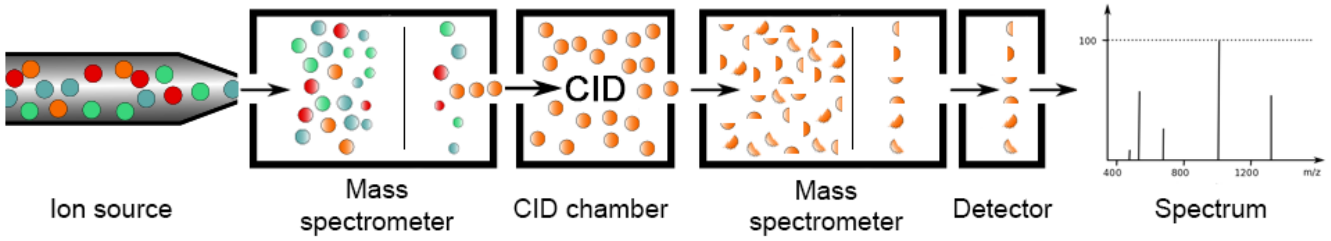


Figure 1.17 – An MS/MS setup with a CID part: an ion source injects a bunch of ions into the first mass spectrometer, which selects ions of interest. These ions undergo CID and their fragments are further sorted by the second mass spectrometer and, at the last stage, a detector identifies the fragments. Using the data obtained an experimental MS spectrum can be built.

A computational approach, described in Subsection 1.4.1, to perform CDS simulations has been developed first by Hase *et al* [44, 45, 46] and further extended by Molina E. S. *et al* [40]. It combined ESI-MS/MS experimental data with CDS at a quantum mechanics / molecular mechanics (QM/MM) level of theory of CID processes to obtain deeper insights into possible fragmentation processes. Among its other advantages, direct dynamics is capable to provide structural information of the fragments, obtained after a collision, thus, it delivers reaction pathways without specifying any reaction coordinates. The theoretical approach requires to specify an optimised geometry structure of a molecule and a set of simulation parameters, such as collision energy, an impact parameter (an ion-projectile distance at the time of collision), a time step and a few others. These parameters are not straightforward to specify —- several preliminary calculations have always to be run to adjust those values, as changing one parameter might affect others. The workflow is thoroughly discussed in Subsection 1.4.2 along with explanations of how to set up CDS.

1.4.1 Theoretical approach for CID calculations

CID processes can be modelled with the help of CDS by calculating an ensemble of simulated trajectories, describing collisions between a projectile ion and inert buffer gas. Relative translational energy is pre-defined and the algorithm samples all possible collision orientations, trying to mimic the ones occurring during CID experiments. To build reliable statistics leading to meaningful conclusions, it is compulsory to run thousands of trajectories. This can be achieved either by utilizing an analytical potential energy function (that introduces unimolecular decomposition paths for given ions) [47] or by direct dynamics [12]. However, the former one can only be used in some limiting cases for which the analytical function can be derived. Thus, in a majority of applications the latter one — a direct dynamics approach — is used by employing MM potentials for ions under consideration. Such CID modelling approach has limitations of the time scale that can be simulated, therefore normally only fast processes are considered. This potentially might prevent from understanding of a full fragmentation process. Nonetheless, it is believed that all crucial fragmentation events occur in the beginning of the process, thus, no important information is lost during CDS [48].

A system under investigation is normally broken down into 2 parts: QM and MM parts. Since with an increasing system size, treating compounds with QM methods becomes computationally expensive, typically only the ion is QM treated, while the buffer gas is simulated with MM potentials. That being said, a computational potential, employed in CDS, consists of two parts:

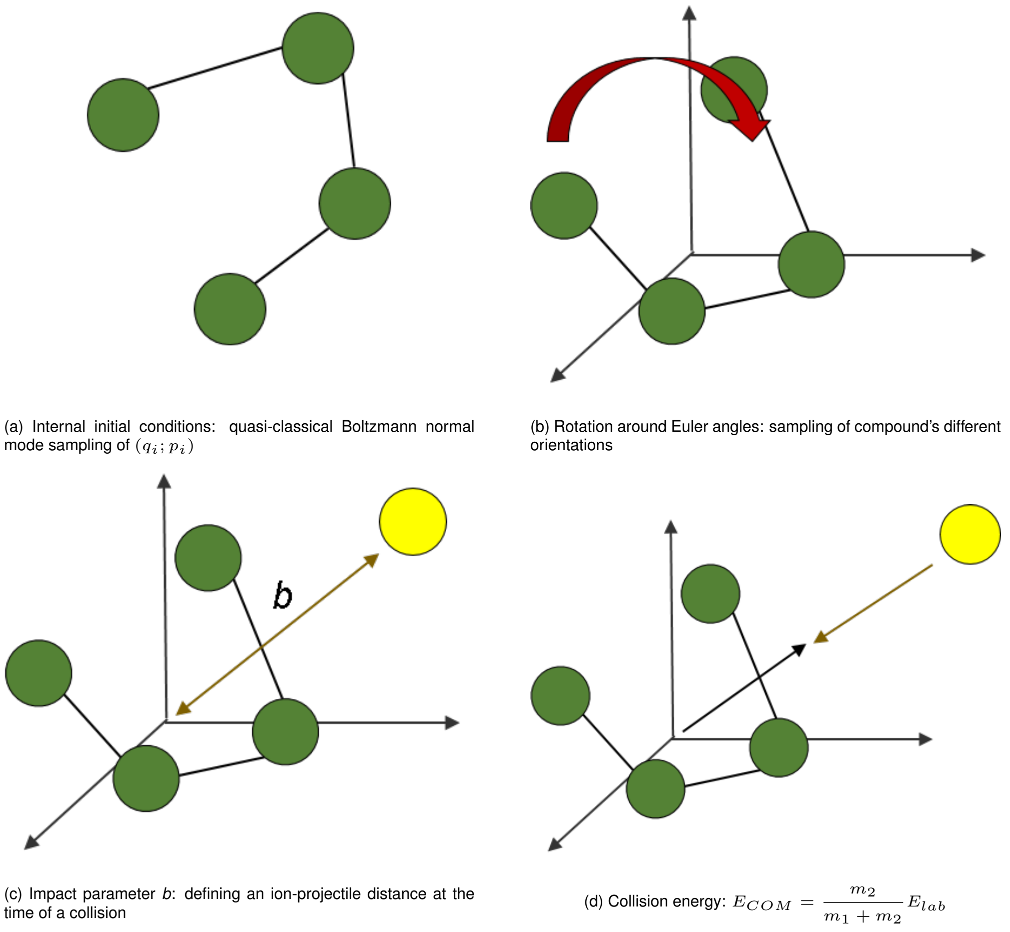
$$V = V_{ion} + V_{gas-ion}, \tag{1.4.1}$$

where V_{ion} is an intramolecular interaction potential, obtained with the help of a QM method, whereas $V_{gas-ion}$ is an analytical potential used to describe ion-projectile intermolecular interaction with MM. The analytical potential, described in [49], was developed by Hase and Meroueh to model CID processes of protonated peptides. As a projectile, one of inert gases, such as *Ne*, *Ar* or N_2, is normally used. Typically, N_2 is chosen due to its lower costs compared to other gases; additionally, this is the gas of choice in IMS/MS experiments, described in Section 1.2. However, CDS performed for the purpose of this project have utilized *Ar*. This is due to the fact, that parameters for the analytical potential, describing ion-projectile intermolecular interaction, were already available for *Ar* and it has been demonstrated by Anderson *et al* [50], that many noble gases behave similarly in CID experiments. Therefore, the second term in Equation (1.4.1) represents the Ar-ion interaction potential and is a sum of two-body elements of the form (as in [49]):

$$V_{Ar-ion} = \sum_i a_i e^{-b_i r_{i,Ar-ion}} - \frac{c_i}{r^9_{i,Ar-ion}}, \tag{1.4.2}$$

where i runs over all ion's atoms, $r_{i,Ar-ion}$ is a distance between each *Ar* atom and each atom of the ion; a_i, b_i, c_i are coefficients obtained by either fitting the analytic potential to an *ab initio* potential, calculated for each atom-atom pair or, alternatively, by looking up the values in literature, where the same atom types were studied.

Having identified interaction potentials, a collection of trajectories was obtained by integrating the classical equations of motion using the velocity Verlet algorithm [51] with a chosen time step, that gave energy conservation for both reactive and non-reactive trajectories. To model numerous initial orientations of the ion at the time of a collision with the projectile, an Euler angles method was used. In addition to this, an impact parameter *b* had to be specified to describe the distance of the closest contact of the ion and the projectile. It can be either fixed to some value or randomly chosen every time during a calculation. Figure 1.18 shows a general workflow of a typical CID simulation, starting with an initial structure preparation finishing up by analysis of fragments.



(a) Internal initial conditions: quasi-classical Boltzmann normal mode sampling of $(q_i; p_i)$

(b) Rotation around Euler angles: sampling of compound's different orientations

(c) Impact parameter *b*: defining an ion-projectile distance at the time of a collision

(d) Collision energy: $E_{COM} = \dfrac{m_2}{m_1 + m_2} E_{lab}$

Figure 1.18 – A scheme of a general workflow of a typical CDS setup: in the beginning, an optimised structure of the ion (a) is rotated around Euler angles (b) following by an assignment of a *b* value — either fixed or randomly chosen (c). After that, a collision between an ion and a projectile with energy E_{COM} takes place (d). In the description of the sub-figure (d): E_{COM} is centre-of-mass energy, m_1 and m_2 are corresponding masses and E_{lab} is energy of the collision in a laboratory reference system.

To calculate trajectories, a combination of a direct dynamics programme VENUS96 [46, 52], linked to an electronic structure code programme MOPAC [53] with a PM6 semi-empirical method [54] enabled, was used. In addition to this, a collection of scripts, written in *bash*, were utilised to perform analysis of simulated trajectories. A typical workflow of how to run CDS is discussed in Subsection 1.4.2.

1.4.2 Setting up and running CDS

CDS cannot be run straight-away without a prior benchmark. A set of parameters, responsible for a flow of simulations, must be pre-defined in order to determine the best accuracy-performance ratio. These parameters have to be chosen once and then can be repeatedly used for simulations with the same or similar structures. A scheme on Figure 1.19 guides through required steps to prepare CDS and corresponding brief explanations are given below:

1. A structure of interest is optimised with MOPAC at the *PM6* level of theory. After that, the geometry is extracted and used in VENUS96 input files. For each atom of the structure, V_{Ar-ion} parameters are found in literature or calculated and are included in the input file.

2. *hinc* — the first parameter to be defined. It is the Cartesian coordinate displacement interval for calculating force constants numerically. *hinc* must be chosen by trial, but in the majority of cases the best value is approximately 0.001Å. The procedure to derive it (and other mandatory parameters) is to choose a particular value of *hinc*, while fixing all others, and to run calculations. Once this is done, a few criteria must be checked to verify choice correctness. If a selected value satisfies conditions — it is set to be a default value for next calculations. If not — a new parameter is chosen and a new set of calculations is run until a satisfactory value is found. For *hinc* to be approved, calculations with it should lead to rotational and translational frequencies being equal to 0, while producing vibrational frequencies as similar as possible to the ones, obtained from the initial MOPAC calculation.

3. A time step and R_{max} are the next parameters to be specified. If the former one is rather self-explaining, R_{max}, on the other hand, is a distance between a projectile and ion's fragments (in Å), at which the trajectory is halted. These parameters are accepted if the percentage of trajectories that pass an energy conservation check is higher than 75% and if one can verify that simulations were stopped after all main fragmentation events have occurred. The first condition is due to the fact, that if many trajectories have energy dissipation, meaning that the energy is not conserved, one may end up running very time demanding calculations, among which at least 25% (more than a quarter, as picked here as a threshold) will be of no real use, therefore computational resources will be wasted. The second condition ensures that the ion has enough time to undergo fragmentation process (while the projectile is flying away).

4. E_{coll} defines collision energy. It is a crucial parameter since it directly influences on a number of reactive trajectories, which are at the focus of these calculations. Increasing E_{coll} leads to transfer of a larger amount of energy to the ion, thus, causing more potential fragmentations to occur. However, abusing this parameter and setting it to abnormally big values may produce unrealistic results (e.g. an ion can simply explode into pieces after a collision), especially if such energies are not observed in experiments. An acceptable

margin of reactive trajectories is 10% of all simulated trajectories that survive the energy conservation check. If a chosen E_{coll} value leads to such output — it will be used by default.

5. Once all these parameters have been identified, a large number of trajectories is run to get statistics. Energy conservation check is performed at the end to filter out all "non-conserved" trajectories. The remaining trajectories are converted to *.xmol* video files and are used for further fragmentation analysis, that provides such information as fragments count, corresponding masses, broken/formed bonds during a fragmentation process, etc.

Running CDS can be very time demanding, depending on computational time available. Therefore choosing right input parameters is important and can potentially save time, while providing high-quality data. Results of CID simulations, performed for the purpose of this PhD project, are presented and discussed in Chapter 3.

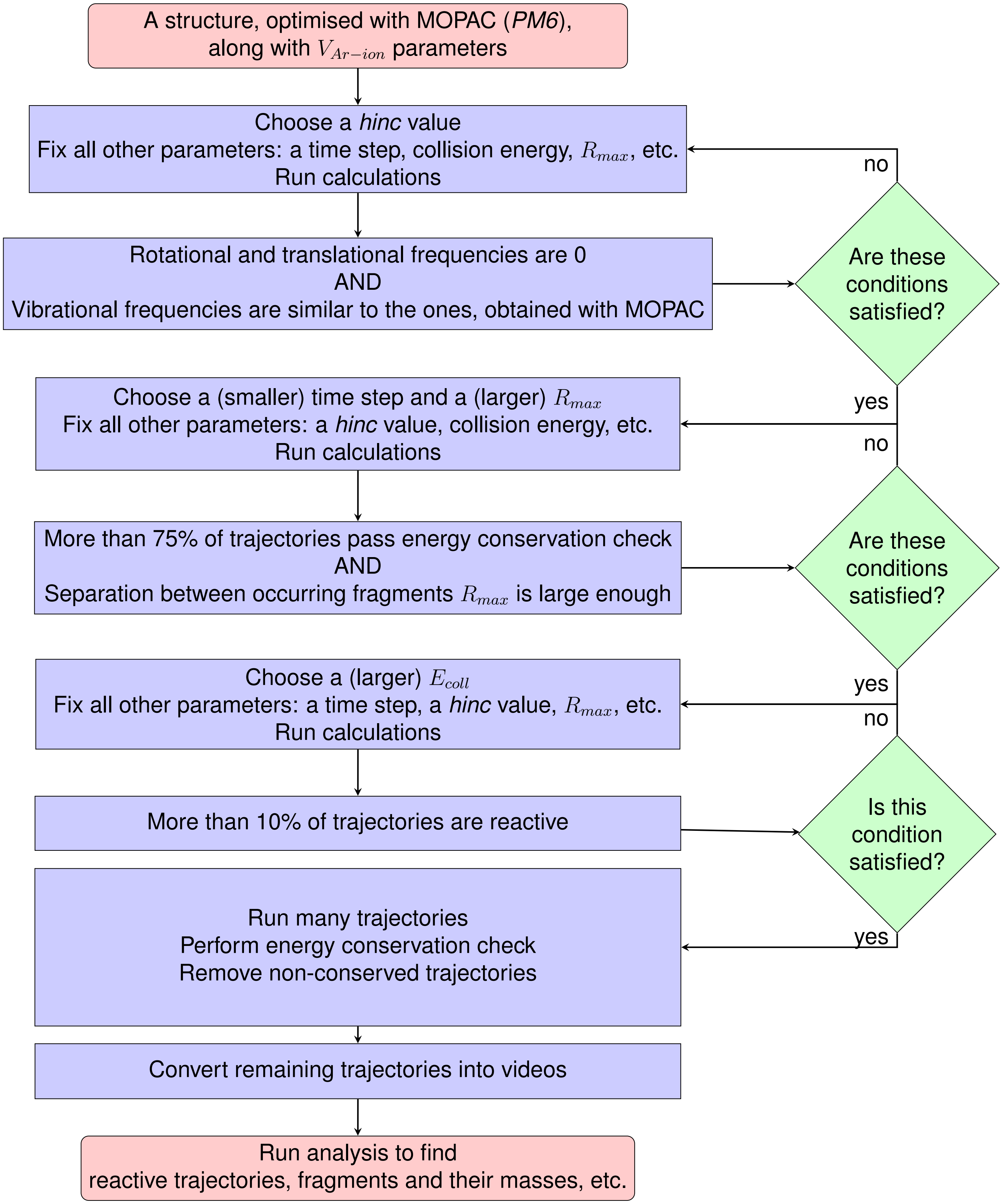

A structure, optimised with MOPAC (*PM6*), along with V_{Ar-ion} parameters
Choose a *hinc* value
Fix all other parameters: a time step, collision energy, R_{max}, etc.
Run calculations
Rotational and translational frequencies are 0
AND
Vibrational frequencies are similar to the ones, obtained with MOPAC
Are these conditions satisfied?
no
yes
Choose a (smaller) time step and a (larger) R_{max}
Fix all other parameters: a *hinc* value, collision energy, etc.
Run calculations
More than 75% of trajectories pass energy conservation check
AND
Separation between occurring fragments R_{max} is large enough
no
Are these conditions satisfied?
yes
Choose a (larger) E_{coll}
Fix all other parameters: a time step, a *hinc* value, R_{max}, etc.
Run calculations
no
More than 10% of trajectories are reactive
Is this condition satisfied?
yes
Run many trajectories
Perform energy conservation check
Remove non-conserved trajectories
Convert remaining trajectories into videos
Run analysis to find
reactive trajectories, fragments and their masses, etc.

Figure 1.19 – A flowchart describing logic behind setting up CDS. Several runs of VENUS96 with various inputs are required in order to identify the most suitable parameters. It is necessary to make sure that calculations take a reasonable amount of time while providing satisfying level of accuracy.

2. Protocol

An extensive computational protocol has been developed by Reading, E. *et al* [11] to perform analytical CCS calculations, utilising a specific list of tools along its execution. The authors compared the results to an experimental output, obtained via the IMS-MS technique, described in Section 1.2. The protocol was run against a set of compounds, for which experimental values had been previously measured in the TW mode by a group of experimentalists. A thorough analysis of theoretical vs experimental results was presented in the same paper [11]. A considerable part of this PhD project has been devoted to the improvement of the protocol and it played a central role in project's further development decision-making.

2.1 Overview

At the initial stage of the project, the protocol required Chemical Table *.mol* files or SMILES *.smi* files as its input. SMILES stands for a Simplified Molecular-Input Line-Entry System and has been developed by Weininger D. in 1988 [55]. It is a chemical notation system, based on a molecular graph theory, whose purpose is to simplify modern chemical information processing. The system introduces a natural and intuitive grammar that makes rigorous structure specification possible. Another strong point of SMILES is its high compatibility for high-speed machine processing. As a result, both chemists and analytical algorithms share the same chemical language that allows development of many highly efficient chemical computer applications able to produce a unique notation, to perform fast database searches and structural information transfer, to build property prediction models, etc.

The protocol, whose simplified workflow is schematically illustrated on Figure 2.1, included the following steps:

1. **Input preparation** — protonation or deprotonation of a compound, is a first step of the protocol. This is important as experiments can be done either with positive or negative ions. Deprotonation is done by manually and sequentially removing a Hydrogen atom from a parent structure, using any convenient chemical structure editor, and saving the resulting output in a separate file. Protonation, on the other hand, is done by means of a Bourne

Again SHell (*bash*) script, which edits a compound's *.smi* file by adding a Hydrogen atom to all Nitrogen and Oxygen atoms and eliminating unrealistic chemical structures in step 4.

2. After the initial preparation, the protocol invokes MOE (**Molecular Operating Environment**), developed by Chemical Computing Group[4], which is a Computational Chemistry toolkit widely used to perform molecular modelling and visualisation tasks. Once a given compound is imported into the software, energy minimisation with the *PFROSST* force-field [56, 57] is performed along with a subsequent *AM1-BCC* semi-empirical partial charges calculation [58, 59]. The next bit is a conformational search done with the LowModeMD algorithm (more details in Subsection 2.1.1) and export of the resulting database into a Structure Database *.sdf* file [60].

3. Subsequently, the output *.sdf* files are converted to a Gaussian input format with the help of a molecular converter **BABEL** [61, 62].

4. Then, quantum refinement of the obtained conformations is performed using a **Gaussian software package** [63]. The level of theory used is an *AM1* semi-empirical method, which is proved to provide satisfying results, compared to other methods, such as DFT or HF [11]. At this point, all calculations with chemically unrealistic structures, obtained by protonating all possible positions of *O* and *N* in the initial stage, crash, and, thus, are removed from the further analysis.

5. The previous step is followed by another instance of **BABEL** converting Gaussian *.out* output files into *.pdb* files, serving as an input for the MOBCAL software.

6. **MOBCAL** [30, 31], introduced in Subsection 1.3.5, calculates CCS using one of the three different techniques: PA, EHS and TM, with the last one being the most refined technique taking into account a force-field, but being very computationally expensive.

7. Finally, the CCS values, obtained with MOBCAL, along with energies from the Gaussian calculations are imported into **Spotfire**[5] —- an advanced analytical tool, where a "Boltzmann average" is calculated providing a final averaged CCS value.

2.1.1 LowModeMD Conformational Search

LowModeMD is a stochastic conformation generation protocol which is based on perturbing an existing conformation along a molecular trajectory using initial atomic velocities with kinetic energies concentrated on low-frequency vibrational modes, followed by energy minimisation. It was developed in 2009 by Labute P. [64] and is extensively used in MOE. This is a critical point as

[4] `http://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm`

[5] `https://www.tibco.com/products/tibco-spotfire`

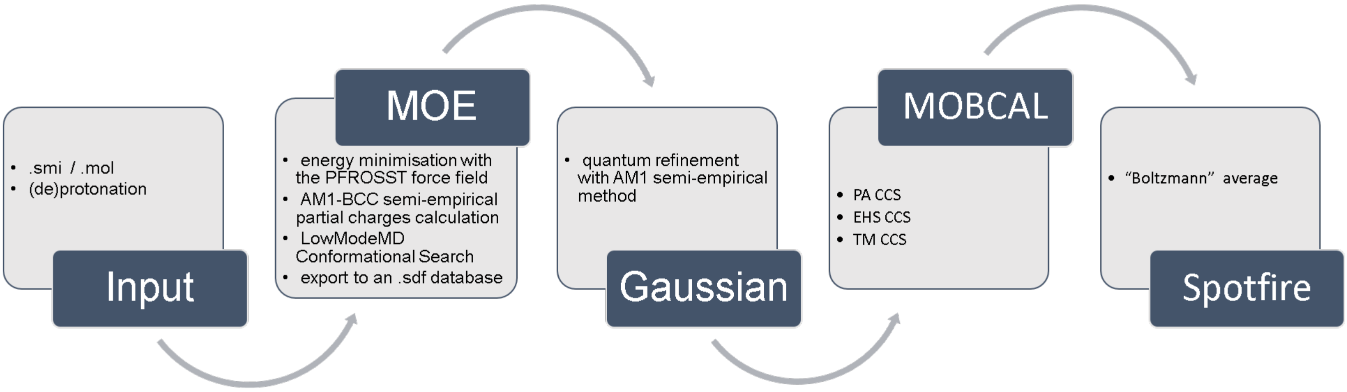

Input
• .smi / .mol
• (de)protonation
MOE
• energy minimisation with the PFROSST force field
• AM1-BCC semi-empirical partial charges calculation
• LowModeMD Conformational Search
• export to an .sdf database
Gaussian
• quantum refinement with AM1 semi-empirical method
MOBCAL
• PA CCS
• EHS CCS
• TM CCS
Spotfire
• "Boltzmann" average

Figure 2.1 – A simplified workflow of the protocol presented in [11]. Each box represents user's interaction with some piece of code (either input files preparation or launching some software package). At this point, presence of a user was crucial since transferring output from one software package to another one had to be done manually.

this adds random uncertainty to the conformational search process that may affect its accuracy and reproducibility. How to minimise the chances of missing a key conformation is one of the objectives of this study. A few parameters can be changed for this algorithm as presented below:

- **Energy window** — defines the upper bound energy difference between a minimum energy conformation and all other conformations found. Conformations with energies that fall below this value will be saved, whereas all other conformations will be rejected. Its default value in the protocol was $2\ ^{kcal}/_{mol}$.
- **Iteration limit (IL)** — sets up how many iterations the software will perform trying to find new conformations. A default value: *10000 cycles*.
- **Rejection limit (RL)** — specifies how many times in a row the algorithm can fail while trying to find a new conformation, before it terminates the conformational search. A default value: *100 cycles*.

The last two quantities have turned out to be of a crucial value as it will be presented and explained why in Section 2.3.

2.2 CCSblackbox script

As can be seen from Figure 2.1, the protocol involved many different tasks that had to be done manually. It is important to understand that this was very time-demanding and required constant user interaction at every step of the protocol, what made it impractical to be used at an industrial scale. Thus, keeping that in mind, an idea of creating and using an advanced script, which incorporates all the mentioned pieces of software in one so-called "black box", is suggested.

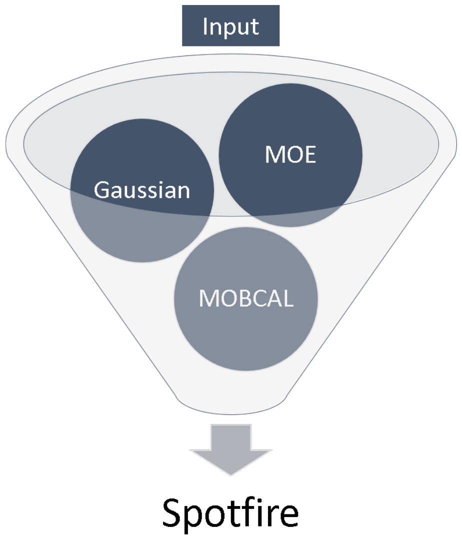


Figure 2.2 – A simple scheme demonstrating the purpose of the *ccsblackbox* script: to incorporate as many steps of the protocol as possible in a black box, where the user does not need to participate; only input must be provided, interaction among the software packages will be done automatically.

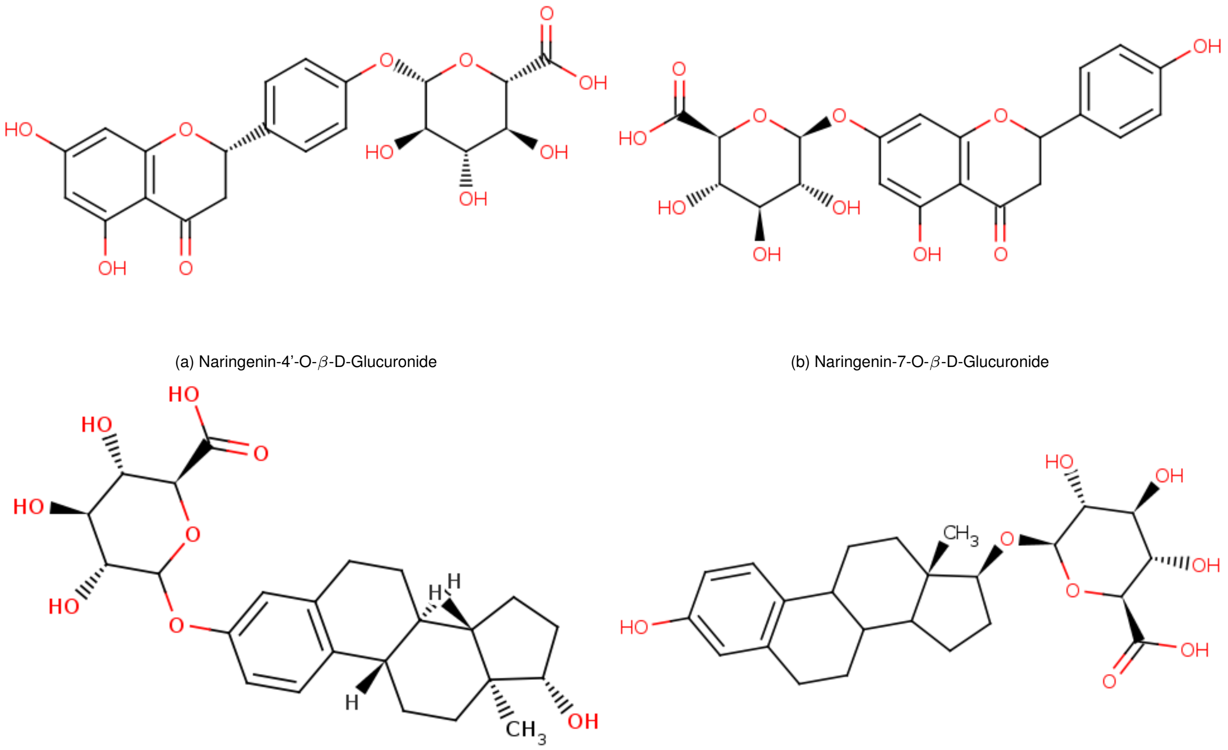
With substantial and careful work the script is now implemented, called for convenience *CCSblackbox*. It is written in *bash* and can be easily updated, if necessary. Figure 2.2 illustrates basic philosophy of the script: a user is only required to specify a desired input and to launch the script; the rest will be done automatically. *CCSblackbox* takes as an input a SMILES *.smi*, Chemistry Table *.mol* or Structure Database *.sdf* file and on an output produces a nicely formatted text *.txt* file with energies from Gaussian calculations and CCS values, that can be subsequently imported into Spotfire. Additionally, the script includes updates to reflect the changes, introduced in the protocol in Subsection 2.3 and Section 2.3.3. The availability of this script allows to perform many calculations in a minimum amount of time limited only by the time needed exclusively for calculations and not for user interaction among software packages.

2.2.1 Scientific Vector Language (SVL)

Some routines of the script are written in the Scientific Vector Language (SVL) [65]. It has been developed by the same company as MOE (Chemical Computing Group) and it permits to avoid the execution of a graphical interface of MOE and to run all necessary commands directly from a command line. Among its responsibilities are the interaction of the *bash* shell environment with MOE, taking care of energy minimisation, partial charge calculation, conformational search with LowModeMD and export to a database. The implementation of the SVL part gives an additional degree of freedom to the protocol as it allows to programme protocol's flow without involving a user.

2.3 Validation

With the help of the *ccsblackbox* script, several trial tests have been made with the purpose of reproducing the earlier published results in [11]. As a test set, four studied compounds — two pairs of isomers — have been chosen (see Figure 2.3): Naringenin-4'-O-β-D-Glucuronide (a), Naringenin-7-O-β-D-Glucuronide (b), β-Estradiol 3-(β-D-glucuronide) (c) and β-Estradiol 17-(β-D-glucuronide) (d).



(a) Naringenin-4'-O-β-D-Glucuronide

(b) Naringenin-7-O-β-D-Glucuronide

(c) β-Estradiol 3-(β-D-glucuronide)

(d) β-Estradiol 17-(β-D-glucuronide)

Figure 2.3 – A test set used for the protocol validation. It included 2 pairs of drug metabolite isomers, taken from [11].

After running the script against the test set using identical parameters, as in the original paper [11], it turns out, that there is some systematic inconsistency between the newly calculated and the published values. Due to the complexity of the protocol the issue had to be tracked down by checking all intermediate steps within the protocol, thus, supplementary analysis was performed. After reassuring that "Boltzmann averages" in Spotfire had been calculated correctly, the CCS calculations in MOBCAL were run with the TM method, which is believed to be the most accurate. Since it did not lead to any further improvement, a higher level of theory (Hartree-Fock) was used for quantum refinement in Gaussian calculations. Nonetheless, the discrepancy between the results was not eliminated. The next attempt was to check the LowModeMD parameters in MOE (described in 2.1.1). Changing the default values of these parameters ($IL = 10000$, $RL = 100$) considerably improves the reproducibility of the published results, underlying the importance of the Iteration and Rejection Limits in the conformational search, and, subsequently, in CCS calculations using the protocol. A large number of calculations was run aiming to find the most suitable pair of parameters reproducing the published results, which at the same time were, to some extent, in accordance with experimental values. It had to be done carefully as larger parameters would heavily affect script's execution time, as the Conformational Search is the most time demanding part of the protocol.

In total, the *ccsblackbox* script was run more than 600 times for the test set. The PA method was considered as a reference method and overall protocol accuracy was judged by the final

CCS values the script had produced. On top of it, the EHS and TM methods were also run and their outputs were used for comparison and further investigation.

Remark 2.3.1. *A theoretical value was considered correct if its PA value was within a 6% range of the experimental result — the expected error in the previous paper [11]. The 6% margin has been introduced to take into account errors in theoretical calculations leading to some small value distribution as well as to handle inaccuracies in experiments.*

2.3.1 Identification of optimal IL and RL parameters

To determine the optimal IL and RL parameters, one compound was chosen from the test set (Naringenin-4'-O-β-D-Glucuronide). Additionally, further in the text only one value might be given to describe these parameters (unless specified explicitly) as they were set to be equal ($IL = RL$), meaning that the Conformational Search will give up trying to find a new conformation, after failing a specified number of cycles in a row, only if the last cycle will be the terminating cycle of the whole search. This setting allows the algorithm to explore the conformational space more rigorously. Therefore calculations were run with one of the following parameters: $IL = 10000$ and $RL = 100$ (default); 10000; 20000; 80000 and 300000 cycles. Table 2.1 summarises some of the data obtained for values of 100, 80000 and 300000. Other values are not included in the summary table for the sake of clarity: only the limits leading to the most evident changes are reported.

Remark 2.3.2. *A theoretical CCS value for Naringenin-4'-O-β-D-Glucuronide is 124 Å.*

All three methods for CCS calculation (PA, EH and TM) have been employed and the script has been run 16 times to have better outcome statistics. By "runs", instances of the executed script with exactly the same input structure and parameters are meant. Identical results cannot be obtained since, as mentioned before, the LowModeMD Conformational Search has a stochastic nature. It is worth noting that the choice of a number 16 as a number of runs is based solely on initial guessing of how many runs can be enough to see if one gets any discrepancy among results from the same runs. Thus, after the 16th run, it was concluded that one may already spot outliers with confidence and that there is no need to go beyond that value (the point about outliers is explained further in this subsection).

The values, written in bold and highlighted with a green colour in the table are within the 6% range with respect to the experimental results. Throughout the simulations, it had been observed that a drastic improvement was achieved already after increasing the RL to the value of 10000. Table 2.1 shows that for the default value of the RL – 100 iterations, out of 16 identical protocol runs, only 5 values by PA and EHS were close to the experimental value, while the rest being

	IL and RL: 100			IL and RL: 80000			IL and RL: 300000		
	PA	EHS	TM	PA	EHS	TM	PA	EHS	TM
Run 1	136,19	147,05	**125,40**	**125,21**	**135,20**	**117,31**	**125,49**	**135,60**	**117,62**
Run 2	140,97	151,67	**128,98**	**125,05**	**134,85**	**117,32**	**125,20**	**135,20**	**117,31**
Run 3	140,84	151,60	**128,92**	**124,86**	**135,05**	**117,76**	**125,21**	**135,29**	**117,85**
Run 4	136,74	147,44	**125,42**	**125,21**	**135,29**	**117,85**	**124,86**	**135,03**	**117,76**
Run 5	140,50	151,15	**128,69**	137,09	147,82	**126,17**	**125,19**	**135,36**	**117,65**
Run 6	136,96	147,68	**125,57**	**129,53**	**139,98**	**121,00**	137,08	147,82	**126,17**
Run 7	**127,04**	**137,96**	**119,92**	**129,78**	**140,25**	**121,12**	134,41	145,26	**124,45**
Run 8	141,26	151,93	**129,36**	**126,31**	**136,13**	**117,63**	**127,83**	**138,25**	**119,65**
Run 9	**127,05**	**137,99**	**119,94**	**126,41**	**136,73**	**119,08**	**126,42**	**136,73**	**119,05**
Run 10	136,78	147,70	**125,78**	**126,39**	**136,68**	**119,04**	**126,41**	**136,73**	**119,08**
Run 11	142,40	152,84	**130,14**	**125,75**	**136,43**	**119,06**	**125,77**	**136,44**	**119,06**
Run 12	140,31	151,01	**128,68**	**126,45**	**136,89**	**119,31**	**125,77**	**136,44**	**119,07**
Run 13	**126,16**	**136,86**	**119,31**	**126,45**	**136,80**	**119,32**	**126,49**	**136,84**	**119,35**
Run 14	141,30	151,80	**129,14**	**126,50**	**136,99**	**119,32**	**124,53**	**133,76**	**115,35**
Run 15	**126,89**	**137,62**	**119,86**	**125,71**	**136,39**	**119,14**	**125,78**	**136,48**	**119,23**
Run 16	**125,34**	**135,33**	**117,18**	**127,59**	**138,48**	**120,40**	136,88	147,63	**125,60**

Table 2.1 – A comparison of theoretical CCS values for Naringenin-4'-O-β-D-Glucuronide obtained with PA, EHS and TM methods for a set of IL and RL parameters. An experimental CCS value is 124 Å. Cells highlighted with bold font and a green colour indicate that a reported CCS lies within a 6% range of the experimental value.

wrong. Increasing the RL to 20000, and then to 80000 cycles, greatly changed the outcome – 15 correct results by PA and EHS and only 1 wrong. And in the case of the largest RL, presented here, 300000 iterations, 13 correct results and 3 wrong have been reported by the PA and EHS methods. Thus, increasing the limits in the LowModeMD Conformational Search helps to improve the reproducibility of the results, however chances of getting misleading values still exist. Calculations with the RL of 300000 cycles normally take about 3-4 days to be completed on a workstation with an Intel i5 processor with 8 Gb RAM, therefore making the protocol unsuitable to be used effectively, since running many instances of it for a dozen of compounds would take weeks. For all discussed cases, the PA and EHS have shown a similar trend in a number of correct CCS values, whereas the TM method has shown a rather consistent pattern, providing all results within the desired 6% range of the experimental value. However, one may notice, that the closest to the experimental values results are provided by the PA method with a higher number of the IL and RL limits. Moreover, it has been observed that there existed some critical values for the two parameters after which no improvement could be obtained (increasing the limits from 80000 cycles to 300000 cycles has not produced any improvement).

These trends can also be seen on Figures 2.4-2.6, which demonstrate a crucial point that a LowModeMD is a stochastic algorithm — one can see oscillations in CCS theoretical values for all the CCS calculation methods. In addition to this, it has been observed that the PA method produces values the closest to the measured ones, while the EHS method tends to overestimate a CCS value. Interesting to note that already at the IL and RL values, larger than 10000, TM and PA CCS results were in a rather good agreement with the experimental ones, thus not requiring time consuming modelling with large conformational search parameters, however, in the case of TM, a constant underestimation from the experimental CCS value suggests that some fundamental element might be missing and needs to be further investigated (e.g., correct potential's parametrisation, see Subsection 2.3.4). Not much could be done in terms of improving the PA and EHS methods as they are rather trivial, but since the PA method has provided satisfying results with a larger number of iteration cycles, it was chosen to be a standard method for the protocol. It is crucial to point out here that this decision is only acceptable if small molecules, interacting with *He* gas, are considered. PA will fail in other instances where compounds of large size and different background gas are used. Therefore to address those cases a more precise technique should be used, however, since this work considers only small molecules (drug metabolites), and speed of calculations was of high importance (due to the protocol's potential application in industry), the PA method was selected as the one, providing the desired properties. Moreover, the leading aim, at this moment, was to eliminate the randomness, present in the protocol, to ensure delivery of reliable results.

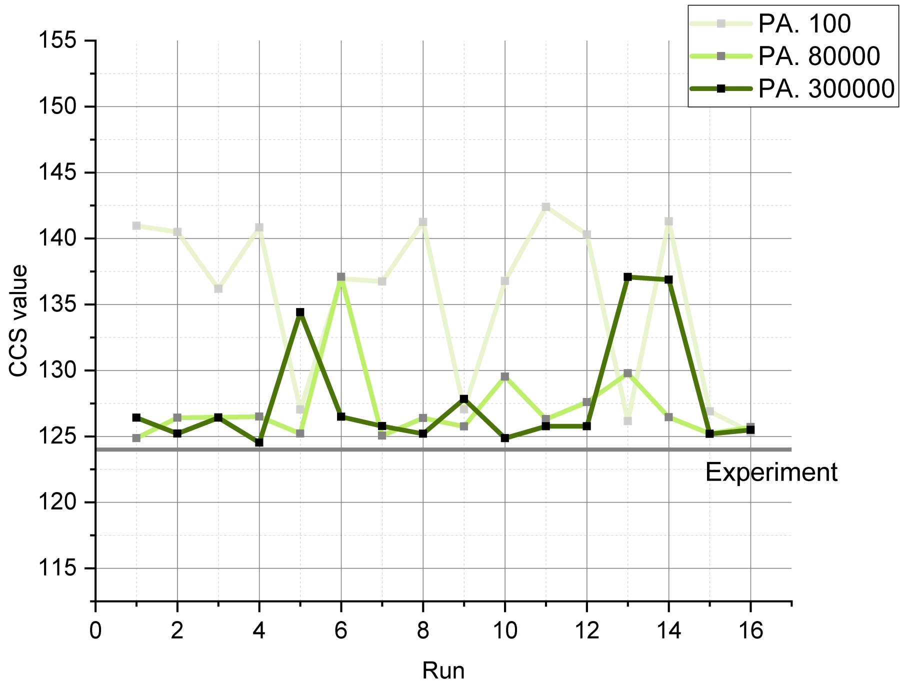


Figure 2.4 – An illustration of the stochastic nature of the LowModeMD Conformational Search algorithm. Three plots represent data, obtained using the PA method, for calculations run with different IL and RL for the Naringenin-4'-O-β-D-Glucuronide compound.

Similar analysis has been performed for the remaining compounds too and the results are

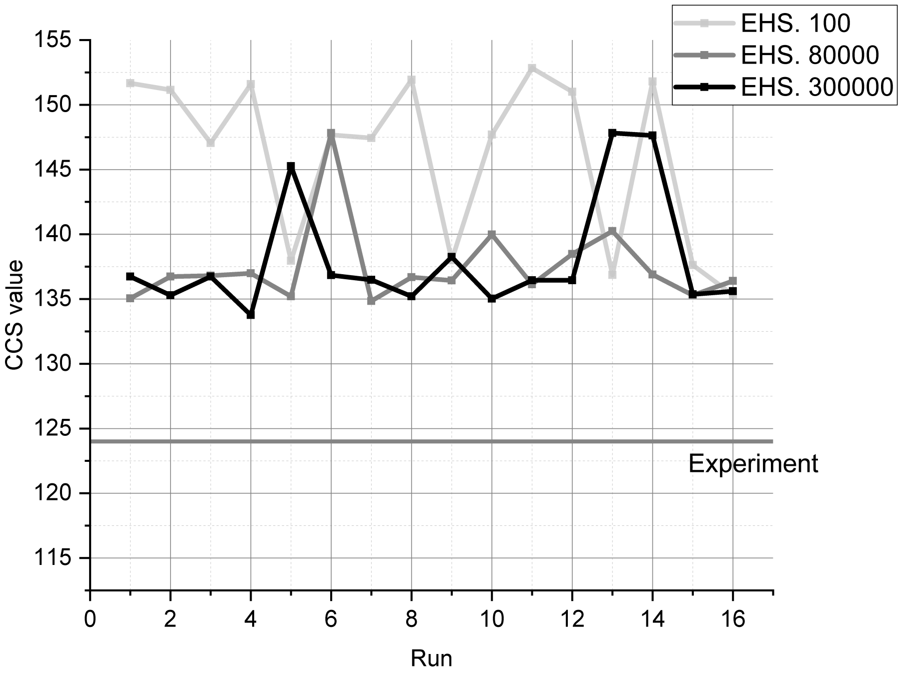


Figure 2.5 – An illustration of the stochastic nature of the LowModeMD Conformational Search algorithm. Three plots represent data, obtained using the EHS method, for calculations run with different IL and RL for the Naringenin-4'-O-β-D-Glucuronide compound.

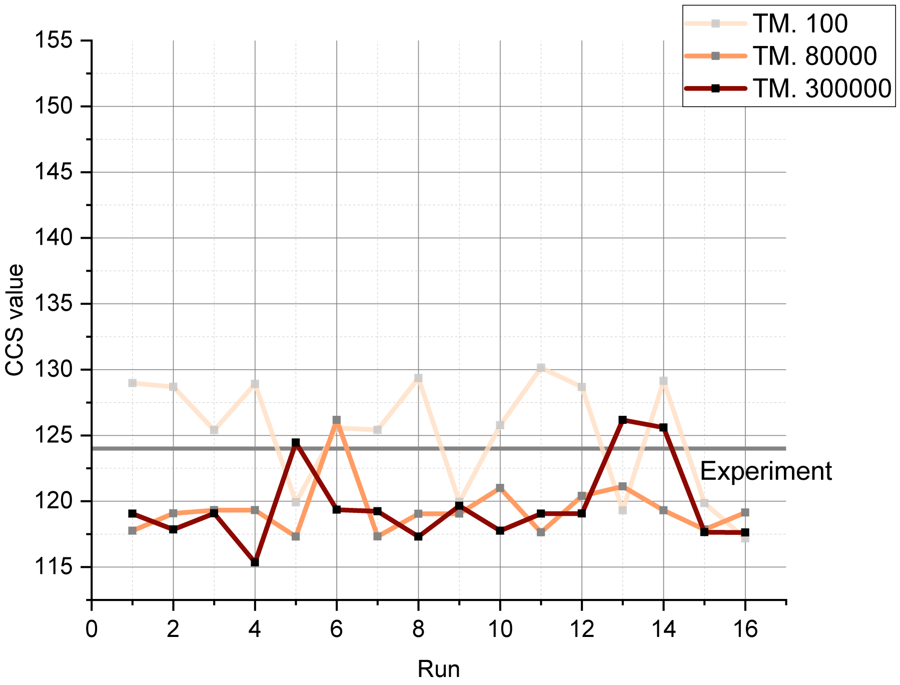


Figure 2.6 – An illustration of the stochastic nature of the LowModeMD Conformational Search algorithm. Three plots represent data, obtained using the TM method, for calculations run with different IL and RL for the Naringenin-4'-O-β-D-Glucuronide compound.

plotted in Figure 2.7 (including some other IL and RL values, not reported in Table 2.1). As can see observed, the value of 50000 cycles is located in between RL values, that provide good results for all metabolites and there is no need to lose in computational time trying to get essentially the same results by increasing the parameters. Therefore, after the validation of the protocol with the *blackbox* script, the standard values for the IL and RL are redefined as a trade-off of accuracy of results and the simulation time required - 50000 iterations. This value is used in all subsequent calculations. The graph also emphasises the necessity to perform as many identical protocol runs as possible to collect meaningful statistics. This is due to the stochastic nature of the LowModeMD Conformational Search algorithm that occasionally might produce conformations, eventually leading to wrong CCS values. Therefore it is necessary to utilise a rigid outlier detection algorithm to identify correct results and to eliminate outliers. This is motivated by the fact, that, ideally, the studied protocol will potentially be used for unambiguous drug metabolite isomers identification in cases, when an experimental CCS value is not known, thus it is crucial to be able to make a proper selection of results among all the data produced by the script. One of such methods (a modified Z-score method) is discussed in Subsection 2.3.2.

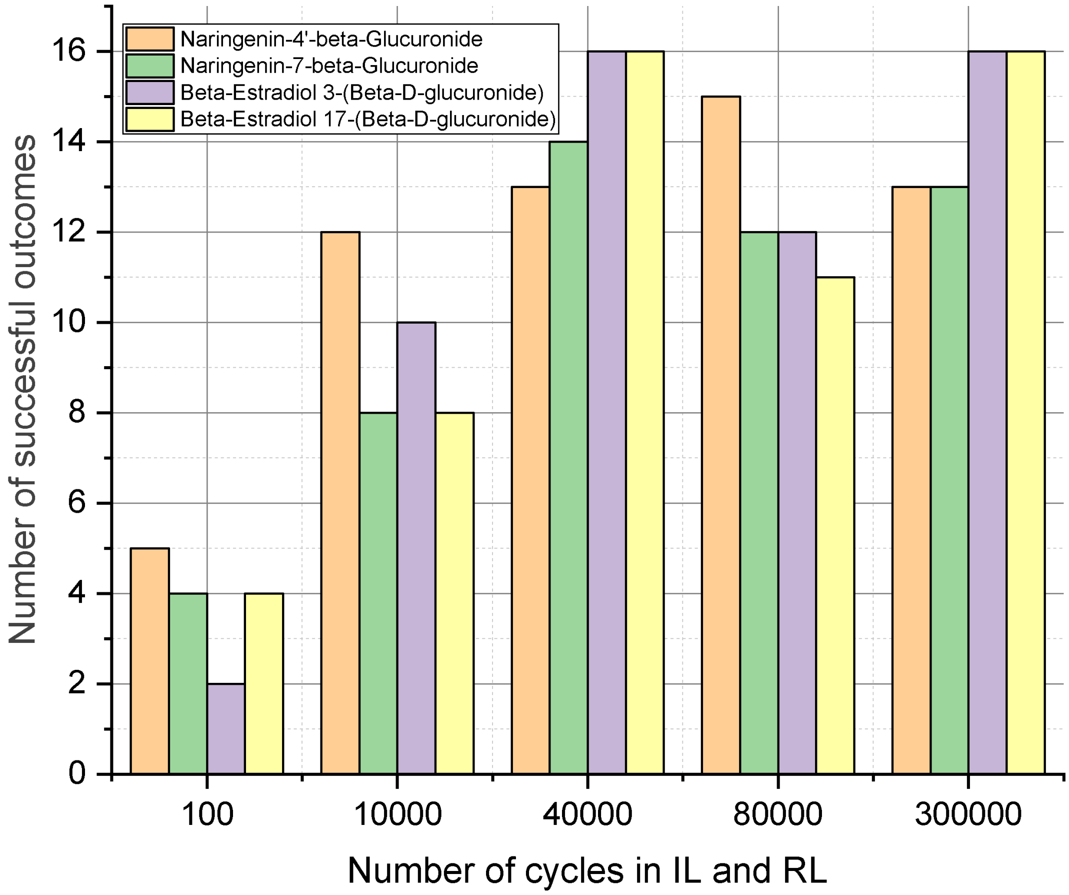


Figure 2.7 – A plot showing a number of successful calculation outcomes vs a number of cycles in IL and RL. The results, taken into account here were, produced with the PA method.

2.3.2 Isomers identification

Having identified a "sweet spot" for IL and RL, calculations aiming to test the protocol's ability to distinguish between isomers were run. The test set consisted of the same compounds as in Subsection 2.3 and the protocol was run with the same parameters, but with fixed values of IL and RL. The PA method was the only one used in MOBCAL as it showed the best performance during the previous evaluations. On the other hand, the calculations were additionally run with He gas being an environment during the LowModeMD Conformational Search, in which an ion was simulated (a standard environment was water). The idea was that since in real experiments the ion is surrounded by some gas (He, N_2, Ar or their mixture), generating conformations in He could potentially lead to better agreement with experimental data.

Taking into account the previous discussion regarding outliers, a simple, but yet efficient algorithm has been implemented in the workflow by means of an ordinary Excel spreadsheet. With its help outliers are identified and the algorithm, employed to locate anomalies, was suggested by Iglewicz, B. and Hoaglin, D. [66] and is presented below:

- Calculate data set's median $\tilde{X}$ (using Excel, Spotfire, etc.).
- Calculate a Median Absolute Deviation (*MAD*) for a data set, defined as:

$$MAD = median(|X_i - \tilde{X}|), \tag{2.3.1}$$

  where X_i is a value from a data set and $\tilde{X}$ is the median of the data. Thus, *MAD* is the median of the absolute deviations from the data's median $\tilde{X}$.

- Calculate a modified Z-score:

$$M_i = \frac{0.6745(X_i - \tilde{X})}{MAD} \tag{2.3.2}$$

- Identify outliers by comparing M_i values to a suggested one. The authors recommend using a Modified Z-Score of greater than 3.5 (as a starting value) as a means to identify possible outliers. However, this value was further tuned with regards to several data sets of interest (CCS values for different compounds), and by comparing a list of outliers, identified by looking at experimental values, and the ones, predicted by the algorithm, the reference modified Z-score value was set to be 2.8.

An example of such outlier detection process is given in Table 2.2. It consists of theoretical CCS values for Naringenin-4'-O-β-D-Glucuronide, performed with $IL = RL = 50000$ and gas being used as an environment. The values, highlighted with a green colour, are considered correct with respect to the experimental value (as explained in Remark 2.3.1, the condition is to be within a 6% range from the measured CCS). Furthermore, grey-coloured cells contain values, that the outlier detection algorithm assumed to be wrong. The table reveals that in the majority of cases, the algorithm correctly marks outsiders, even though some values (2 in the case of the discussed calculations — 131,81Å and 133,29Å) are ignored. Nevertheless, the outlier test successfully removes the values, that otherwise would be the main negative contributors to a final CCS value. Such filtering was performed for all calculations to improve the accuracy of the results.

CCS PA	Diff from median PA	Mod Z-score PA
127,59	0,00	0,00
127,58	-0,01	0,00
126,41	-1,18	0,52
125,73	-1,86	0,83
127,09	-0,50	0,22
127,1	-0,49	0,22
131,81	4,22	1,88
130,06	2,47	1,10
133,29	5,70	2,54
134,56	6,97	**3,11**
136,28	8,69	**3,87**
127,52	-0,07	0,03
136,28	8,69	**3,87**
127,1	-0,49	0,22
126,54	-1,05	0,47
136,28	8,69	**3,87**
Median PA	127,59	
MAD PA	1,52	
Result PA	128,15	
Experiment	124	

Table 2.2 – Outlier detection algorithm applied to the results of calculations with the PA method for Naringenin-4'-O-β-D-Glucuronide. $IL = RL = 50000$, environment - gas. Outliers, identified with the Modified Z-score method, are in bold and highlighted with a grey colour. Green colour signifies theoretical results, that are within a 6% range from the experimental value.

The results of the calculations for isomer identification can be seen in Table 2.3. It illustrates data for two pairs of "unknown" isomers (in quotes since the compounds were known; the idea was to see if it is possible to correctly identify them without prior knowledge of experimental values). For the purpose of the analysis, data from measurements are shown in the "Experiment" row. The outliers are identified and highlighted in bold font and a grey colour. As one can see on Table 2.3, the outlier detection algorithm does a considerably good job at identifying wrong results. By utilising it, it is possible to eliminate values, that would ultimately negatively affect the final result — an arithmetic average of all values, which in this case will be taken only over values, that survived the outlier test. It is necessary to note however, that if all simulation results for a particular compound tend to be over- or underestimated by the protocol, the algorithm will not be able to extract correct values as a core it relies on (most frequent CCS values), will also be shifted. This is what happened in the case of Isomer 1-2, where the test removed values closer to the experimental ones, rather than those, that would be removed manually if the experimental value is known. However, at this point, this is an issue with the protocol itself and not the outlier identification algorithm, as there are no just a few outliers, caused by the stochastic nature of the conformational search algorithm, but all values are shifted, therefore, this must be addressed from the point of rethinking the protocol and finding out what can be a possible issue. This point can serve as a basis for future research. As a result of this shift, identification of Isomer 2-1 (β-Estradiol 3-(β-D-glucuronide)) and Isomer 2-2 (β-Estradiol 17-(β-D-glucuronide)) becomes tricky, as the calculated value for Isomer 2-1 is larger than the corresponding value for Isomer 2-2, whereas experiments suggest the opposite. This might be due to some intrinsic property of the compound that somehow was not taken into account by the protocol. Alternatively, more efficient CCS calculation methods, mentioned in 1.3.5, should be considered to be used, as it can happen that this is simply the limit of accuracy for the PA method. On the other hand, both Isomers 1-1 and 1-2 were distinguished and are Naringenin-4'-O-β-D-Glucuronide and Naringenin-7-O-β-D-Glucuronide, respectively. Overall, these results coincide with the ones, reported by [9] and confirm that still there is space for further investigation of the existing protocol.

2.3.3 Extended protocol

To make a step forward in the improvement of the current protocol, rethinking of the existing workflow to assess possible ways of refining it, should be considered. Therefore, each step of the routine has been examined and a few areas for further improvement were identified. One of them is to make the TM method to be the main approach for theoretical CCS calculation. Additionally, calculation of the atomic partial charges to be later used for the CCS calculation with the TM method is also suggested, and, furthermore, identification of atom types in the resulting structures and introduction of new Lennard-Jones parameters should be implemented. Finally, IM-MS experiments were run using N_2 as a background gas, whereas the theoretical protocol assumed He as the one. This would have to be taken into account by switching the interacting gas from *He* to N_2 or, alternatively, by introducing a correction coefficient derived from the comparison of the experimental and the theoretical values.

To summarise, the protocol still has areas for further testing and improvement and it possibly can be extended with the proposed steps upon successful implementation. All these changes can be easily incorporated into the script by introducing a few new pieces of software, such as

	Isomer 1-1	Isomer 1-2	Isomer 2-1	Isomer 2-2
Run 1	127,59	137,62	144,07	138,33
Run 2	127,58	138,12	**141,02**	140,63
Run 3	126,41	136,48	144,08	141,45
Run 4	125,73	136,27	**142,13**	141,52
Run 5	127,09	137,68	143,92	141,86
Run 6	127,1	136,63	144,08	138,84
Run 7	131,81	138,02	144,07	140,77
Run 8	130,06	137,90	144,07	139,15
Run9	133,29	137,68	**142,15**	140,19
Run 10	**134,56**	138,12	**141,08**	140,68
Run 11	**136,28**	137,68	143,75	138,45
Run 12	127,52	136,58	**144,80**	137,30
Run 13	**136,28**	137,31	**142,15**	**124,18**
Run 14	127,10	139,33	144,08	141,85
Run 15	126,54	136,63	**142,14**	141,85
Run 16	**136,28**	138,02	143,83	138,86
Experiment	124,00	132,70	136,70	140,30
Theory	128,15	137,50	143,99	140,12

Table 2.3 – A comparison of theoretical CCS values for the test set, obtained with the PA method. The purpose is to try to identify which compounds are named as Isomer 1-1, Isomer 1-2, etc. Values highlighted with bold font and a grey colour indicate that a reported CCS is an outlier. The calculations were run using *He* as an environment and $IL = RL = 50000$.

ANTECHAMBER [67]. Apart from that, an independent version of N_2 in-house MOBCAL can be developed, or more preferably, other more efficient pieces of software should be utilised.

2.3.4 Lennard-Jones parameters in MOBCAL

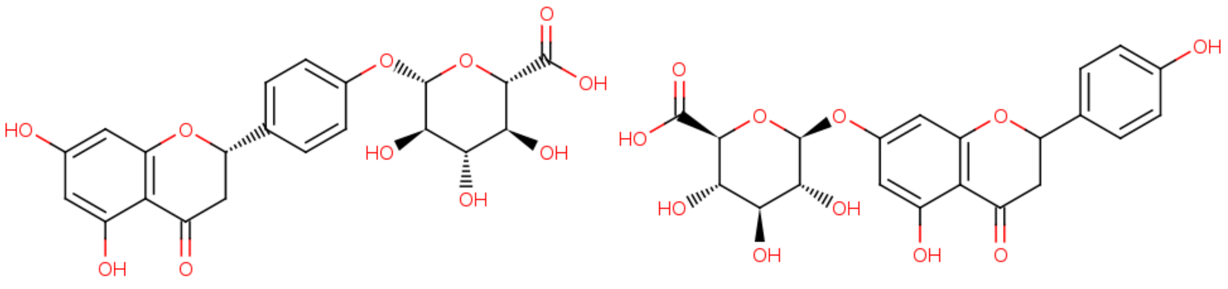
As it has been mentioned before, the potential for the TM method, formerly employed in MOBCAL used in this project, had a simple Lennard-Jones form, where the LJ potential parameters, were defined in the source code of MOBCAL for most of the atoms. Thus, for any Hydrogen atom there was one set of epsilon/sigma values, for any Oxygen atom there was also one set of these values, etc. However, if one looks at popular force fields (e.g. Generalised AMBER Force Field, GAFF, [68]) to see how atomic parameters are defined, it will be discovered that there is an entire range of atomic types present. The definitions and notations might vary depending on the force field, but the main message is clear: by treating all atoms like this, by reducing the complexity to just one type, the accuracy is artificially worsened. So, to take this argument into account, it is possible to take advantage of ANTECHAMBER from the AMBERtools package, mentioned in Subsection 2.3.3. Among its all features, it can identify atomic types in a structure and write them in a file. Having this, one can update the source code of MOBCAL with the atomic parameters available in the GAFF, so that MOBCAL would be able to distinguish atoms of different types and assign parameters appropriately.

3. Chemical Dynamics Simulations

Having thoroughly studied possible ways of improving the theoretical CCS calculation workflow, it was of interest to have a closer look at the other data available from experiments, particularly from MS spectra. The goal was to see if one can gain any useful insights into the collision process by running CDS and if it was possible to identify drug metabolite isomers, for which the computational protocol has failed, with the help of the CDS approach. As in was explained in Subsection 1.4.1 and Subsection 1.4.2, CDS require thorough input parameters preparation, thus, considerable time was spent preparing a satisfying input that would guarantee plausible results.

3.1 Modelling

To be consistent, the same metabolites, used in the previous studies for the CCS calculation protocol improvement, were chosen to be target structures for the dynamics simulations: Naringenin-4'-O-β-D-Glucuronide and Naringenin-7-O-β-D-glucuronide (Figure 3.1). Their experimental MS spectra can be seen on Figures 3.2 and 3.3, respectively (taken from METLIN



(a) Naringenin-4'-O-β-D-Glucuronide

(b) Naringenin-7-O-β-D-glucuronide

Figure 3.1 – Optimised structures of studied isomers.

MS/MS Metabolite Database[6]). As can be seen, the two MS spectra do have distinctive picks, and, therefore can be distinguished.

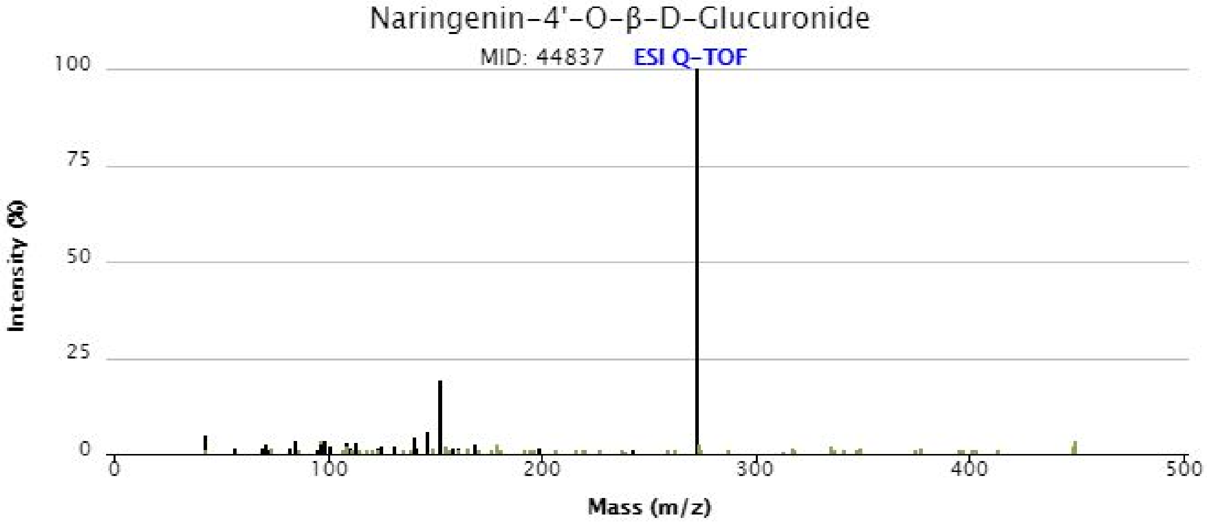


Figure 3.2 – Naringenin-4'-O-β-D-Glucuronide experimental MS spectrum at $20V$, obtained with an ESI technique.

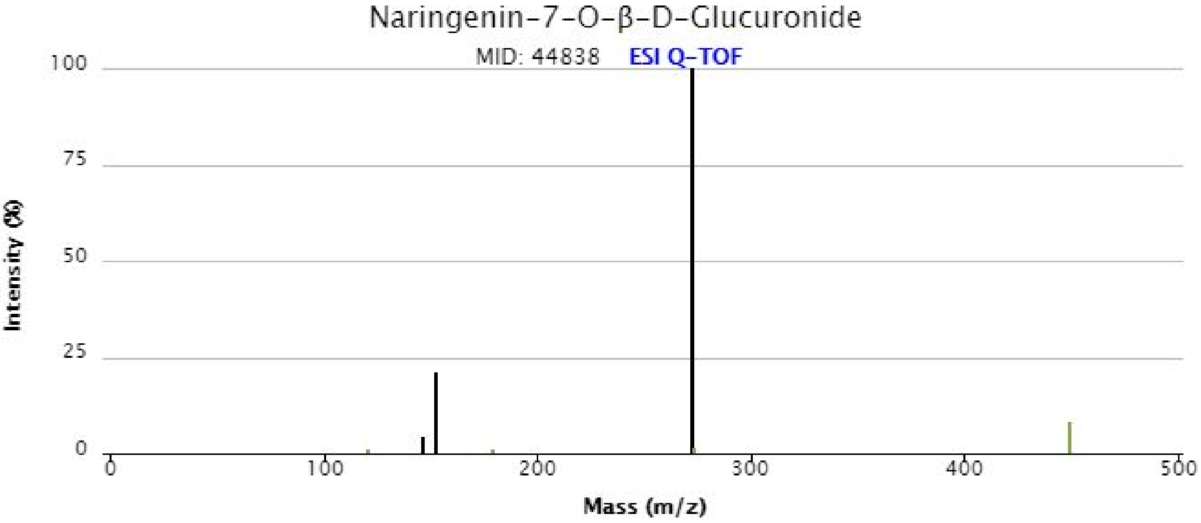


Figure 3.3 – Naringenin-7-O-β-D-Glucuronide experimental MS spectrum at $20V$, obtained with an ESI technique.

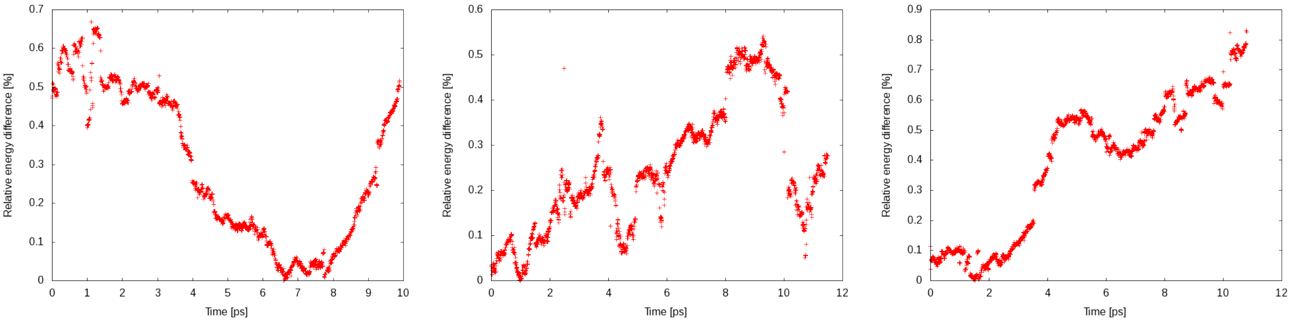
A crucial point while running CID simulations is to establish a balanced time step value, as explained in 1.4.2, since it largely affects simulated time and quality of the results as well as real time the calculations will take to be run on a cluster. A potential, used in the simulations to model an ion-projectile interaction, is a sum of two terms – an ion potential, calculated with a *PM7* semi-empirical method and an ion-projectile potential, which has been explained in Subsection 1.4.1. As it has been mentioned before, modelling was done by coupling MOPAC [53], that was

[6] `https://metlin.scripps.edu/landing_page.php?pgcontent=advanced_search`

responsible for the electronic structure calculation, with VENUS code [46, 52] (a direct dynamics program, [46]). Ar was used as a projectile, even though He or N_2 are mainly used in experiments, however, Ar has been proved to be suitable and more efficient for these calculations, providing reliable results.

3.2 Energy conservation check

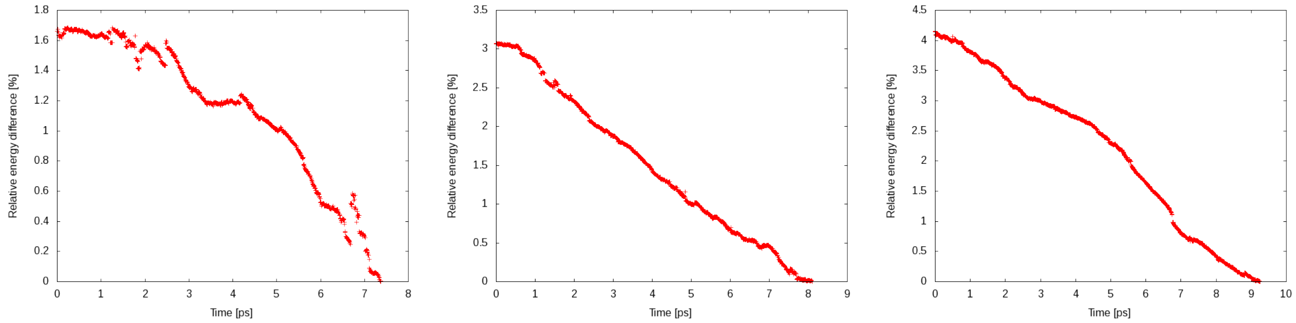
Having run the dynamics, it is necessary to perform energy conservation check. It has been done with the help of in-house scripts. The margin was chosen to be 1% variation, and, thus, those trajectories, which did not fall within the specified range, were considered as outliers and were removed from the further analysis, whereas for the remaining ones, videos were created. Figures 3.4 and 3.5 show a few examples of energy conservation plots for Naringenin-4'-O-β-D-Glucuronide with both — successful and unsuccessful outcomes, run with $E_{coll} = 400{}^{kcal}/_{mol}$. Typically, depending on a set of parameters used, up to 20$\%$ of all trajectories fail to pass the check, therefore one needs to take this estimate into account when a specific number of correct trajectories is required.



(a) Trajectory 1 (b) Trajectory 2 (c) Trajectory 3

Figure 3.4 – Energy conservation plots for a few trajectories for Naringenin-4'-O-β-D-Glucuronide, that passed the check. $E_{coll} = 400\ {}^{kcal}/_{mol}$.

In addition to this, only about 10$\%$ of the conserved trajectories will be reactive, meaning that modelled collisions within those trajectories will lead to compound's fragmentation (that ultimately allows to build a theoretical MS spectra). Thus, normally a large number of simulated trajectories is necessary to build meaningful statistics. Reactivity of trajectories can be increased by using a higher theoretical E_{coll} value, however, one does not want to go too far to unrealistic energies since a produced spectrum will be compared to an experimental one, thus both E_{coll} values must be in agreement. So far, while in search for the most optimal parameters to run calculations for the Naringenin isomers, a different number of trajectories has been first run for Naringenin-4'-O-β-D-Glucuronide for a set of energies (see Table 3.1). An optimal value has been chosen to be $E_{coll} = 400\ {}^{kcal}/_{mol}$ as it provides sufficient reactivity among tested energies and it was further



(a) Trajectory 4

(b) Trajectory 5

(c) Trajectory 6

Figure 3.5 – Energy conservation plots for a few trajectories for Naringenin-4'-O-β-D-Glucuronide, that did not pass the check. $E_{coll} = 400\ {}^{kcal}/_{mol}$.

used for calculations involving Naringenin-7-O-β-D-glucuronide.

	Energy, ${}^{kcal}/_{mol}$				
	250	300	350	400	450
total number	100	200	200	1300	1000
not conserved, %	12	9	8	19	21
reactive, %	0	4	5	8	10

Table 3.1 – A summary of a number of trajectories run for a set of energies for Naringenin-4'-O-β-D-glucuronide, indicating their successful energy conservation check rates and reactivity rates.

3.3 Fragmentation analysis

Created videos (*.xmol* files) were important not only to study fragmentation pathways, but also necessary for a subsequent analysis of the trajectories. Within the analysis, resulting fragments can be identified along with getting some other information, like broken/formed bonds after the collision, fragments masses, etc. Finally, by following theoretical fragmentation pathways, it is possible to identify charged fragments and to construct an MS spectrum. As of now, a thorough analysis of the trajectories is in the progress and theoretical MS spectra are still to be built, but preliminary results indicate that potentially isomers can be distinguished. Figure 3.6 depicts an MS spectrum that contains all fragments, observed during the simulations. That is, not only charged fragments, as they are the ones experimentally registered, but all other ones, produced after the collision. The spectra indicates that there are distinctive peaks, different for both Naringenin drug metabolites, that eventually may serve to identify the isomers. Once full analysis will be completed, corresponding theoretical MS spectra will be produced clarifying if such methodology for isomers identification can be used and/or further extended. Additionally, having more simulated trajectories will lead to more reliable statistics.

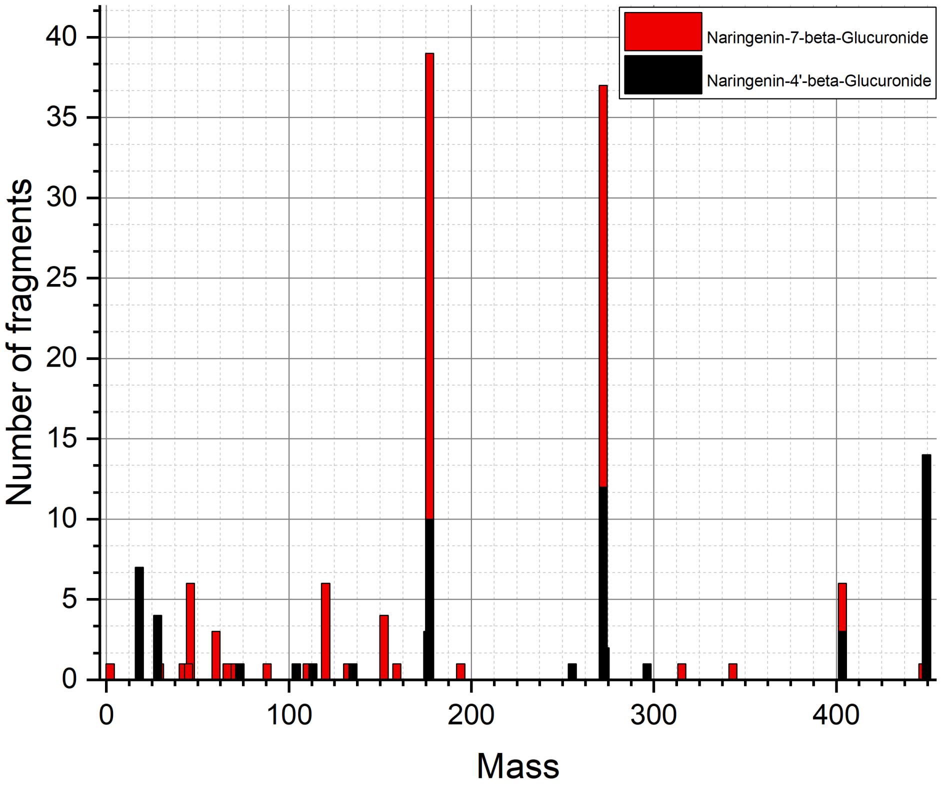


Figure 3.6 – A plot depicting theoretical spectra for both Naringenin-4'-O-β-D-Glucuronide (black peaks) and Naringenin-7-O-β-D-Glucuronide (red peaks) with all fragments' peaks shown.

Conclusões

Neste projeto desenvolveram-se metodologias computacionais que permitem a identificação da estrutura molecular de pequenos fragmentos resultantes de processos de colisão. Em particular, introduziram-se alterações no algoritmo de um protocolo para o cálculo das secções de choque de colisões de metabolitos de modo a otimizar a sua implementação e utilização. Simultaneamente, o trabalho realça o procedimento desenvolvido de parametrização e preparação das simulações de Dinâmica de Colisão e apresenta os resultados desses cálculos.

O desenvolvimento do algoritmo na primeira parte do projeto melhorou muito a interação entre utilizador e protocolo, diminuindo também o tempo envolvido nos vários passos do processo de análise. Um aspeto muito importante a realçar é o facto deste melhoramento do procedimento automático do protocolo permitir o teste de parâmetros de simulação de um modo mais completo e rigoroso, uma vez que minimiza eventuais erros humanos e a dependência de utilização de novas versões de software comercial. Além disso, o algoritmo desenvolvido neste trabalho foi usado para reproduzir resultados previamente obtidos por Reading, E. *et al*, Anal. Chem., 2016, 88 (4).

A existência de alguma inconsistência entre cálculos mais recentes e resultados previamente obtidos mostraram que, devido à sua natureza estocástica, o algoritmo de procura conformacional LowModeMD não se mostrava eficiente na procura de conformações levando, consequentemnete, a previsões de CCS incorretas. Os parâmetros de simulação que se mostraram mais determinantes em LowModeMD foram os limites de rejeição e de iteração. Assim, correram-se centenas de cálculos variando os valores deste dois parâmetros com o objetivo de identificar os valores otimizados que conduzem a resultados CCS em boa concordância com dados experimentais. Concluiu-se que os valores seriam de 50000 ciclos. Testes posteriores confirmaram que estes parâmetros eram a causa da discrepância e, além disso, mostraram a importância de se aumentar a amostra com mais corridas do algoritmo CCS de modo a melhorar as estatísticas dos resultados. Isto é devido ao facto de que, apesar da procura conformacional ter melhorado com a introdução dos novos parâmetros, o algoritmo LowModeMD continua a ter uma natureza estocástica pelo que existe sempre o risco de não se explorar exaustivamente o espaço conformacional.

Após a consolidação do protocolo com os novos parâmetros de simulação, na segunda parte do trabalho passou-se às simulações por Dinâmica de Colisão. Dois isómeros foram

inicialmente escolhidos para análise: Naringenina-4'-O-β-D- Glucuronídeo e Naringenina-7-O-β-D-glicuronídeo, com o objetivo de simular os seus processos de colisão numa matriz de gás e reproduzir os resultados experimentais de espectroscopia de massa. Testes completos de benchmarking levaram à identificação dos valores dos parâmetros mais apropriados para os cálculos VENUS96 e milhares de trajetórias foram executadas para obter estatísticas. A energia de colisão foi definida como sendo $E_{coll} = 400^{kcal}/_{mol}$, pois estava próxima do valor experimental e, além disso, forneceu uma percentagem suficiente de trajetórias reativas. Estas trajetórias serviram de base para a análise de fragmentação, realizada com a ajuda de bash scripts. Os resultados revelam que, de fato, é possível distinguir os isômeros comparando seus espectros teóricos de MS, graças à formação de vários picos distintos. Além disso, as simulações mostraram que, apesar de haver um grande conjunto de fragmentos idênticos, a Naringenina-4'-O-β-D-Glucuronídeo ocasionalmente apresenta diferentes canais de fragmentação da Naringenina-7-O-β-D-glicuronídeo, o que torna viável detetar diferenças. Por fim, uma menor percentagem de trajetórias reativas foi observada para a Naringenina-4'-O-β-D-Glucuronídeo, obrigando a mais cálculos efetuados com este isómero para construir estatísticas apropriadas.

Conclusiones

Este trabajo presenta un enfoque combinado para la identificación estructural de pequeñas moléculas y el análisis de procesos de colisión, que involucran eventos de fragmentación. En particular, propone una forma mejorada de implementar y utilizar un protocolo computacional, desarrollado para calcular las secciones transversales de colisión de los metabolitos de los medicamentos, con la ayuda de un script de shell. Al mismo tiempo, el manuscrito guía a través de un procedimiento de configuración de Simulaciones de dinámica de colisión y muestra los resultados de dichos cálculos, aplicados al proyecto actual.

El desarrollo del script en la primera parte del trabajo, ha mejorado enormemente la interacción del usuario con el protocolo y ha reducido el tiempo necesario para realizar todos los pasos involucrados. Más importante aún, tener una forma automatizada de ejecutar el protocolo dio la oportunidad de probar varios parámetros de simulación de una manera más extensa y rigurosa, ya que se eliminaron los posibles errores humanos y demoras para lanzar el siguiente paquete de software. Además, el guión desarrollado se ha utilizado en este trabajo para reproducir los resultados publicados anteriormente por Reading, E. *et al*, Anal. Chem., 2016, 88 (4).

Después de que numerosos cálculos han revelado cierta incoherencia entre los resultados recién obtenidos y los obtenidos anteriormente, ha resultado que, debido a su naturaleza estocástica, un algoritmo de búsqueda conformacional LowModeMD encuentra conformaciones, que llevan a predicciones de CCS incorrectas. Los parámetros de simulación, que jugaron un papel importante en LowModeMD, fueron los límites de rechazo e iteración. Por lo tanto, cientos de cálculos se han ejecutado con diferentes valores de Límites de rechazo e iteración con el objetivo de identificar un conjunto de estos parámetros, dando como resultado valores de CCS en buena concordancia con los datos experimentales. Se acordó que dichos parámetros fueran iguales a 50000 ciclos. Pruebas posteriores de estos valores han confirmado la hipótesis de que ellos fueron la causa de la discrepancia encontrada y, además, han mostrado la importancia de ejecutar al menos una docena de instancias de protocolo para generar mejores estadísticas a partir de los valores de CCS resultantes. Esto se debe al hecho de que, a pesar de que la búsqueda conformacional con los nuevos parámetros ha funcionado mucho mejor que antes, LowModeMD sigue siendo un algoritmo estocástico, por lo tanto, uno siempre puede terminar por casualidad con resultados que no han explorado completamente el espacio conformacional.

Después de verificar la confiabilidad del protocolo con los nuevos parámetros de simulación, en la segunda parte del proyecto, se cambió el enfoque para realizar simulaciones de dinámica de colisión. Se eligieron dos isómeros para el análisis: Naringenin-4'-O-β-D-Glucuronide y Naringenin-7-O-β-D-glucuronide con el objetivo de modelar sus procesos de colisión con el gas de fondo y reconstruir los experimentos experimentales correspondientes. Espectrometría de masas (MS). Los últimos podrían potencialmente utilizarse para identificar teóricamente los compuestos. El benchmarking ha llevado a la identificación de los parámetros de entrada más apropiados para los cálculos de VENUS96 y se ejecutaron miles de trayectorias para obtener estadísticas. La energía de colisión de trabajo se definió como $E_{coll} = 400^{kcal}/_{mol}$, ya que estaba cerca del valor experimental y, lo que es más importante, proporcionaba un porcentaje suficiente de trayectorias reactivas. Las trayectorias reactivas obtenidas sirvieron como base para el análisis de fragmentación, realizadas con la ayuda de scripts de bash. Los resultados revelan que, de hecho, es posible distinguir los isómeros comparando sus espectros teóricos de MS gracias a varias selecciones distintivas. Además de esto, las simulaciones mostraron que, si bien tienen un gran conjunto de fragmentos idénticos, Naringenin-4'-O-β-D-Glucuronide ocasionalmente experimenta diferentes canales de fragmentación de Naringenin-7-O-β-D-glucuronide, lo que hace posible detectar una diferencia. Además, se observó un porcentaje menor de trayectorias reactivas para Naringenin-4'-O-β-D-Glucuronide, así que se realizaron más cálculos con este isómero para elaborar estadísticas apropiadas.

Conclusions

This work presents a combined approach to small molecule structural identification and analysis of collision processes, involving fragmentation events. It introduces an enhanced way of implementing and utilising a computational protocol, developed to calculate collisional cross sections of drug metabolites, with the help of a shell script. At the same time, the manuscript guides through a procedure of setting up Collision Dynamics Simulations and shows the results of such calculations, applied to the current project.

The development of the script in the first part of the work, has greatly improved user interaction with the protocol as well as decreased the time, necessary to perform all the steps involved. More importantly, having an automated way of running the protocol gave an opportunity to test various simulation parameters in a more extensive and rigorous way, since possible human errors and delays to launch the next software package were eliminated. Furthermore, the developed script has been used in this work to reproduce earlier published results by Reading, E. *et al*, Anal. Chem., 2016, 88 (4).

After numerous calculations have revealed some inconsistency between newly obtained and already available results, it has turned out, that due to its stochastic nature, a LowModeMD conformational search algorithm occasionally finds conformations, leading to incorrect CCS predictions. The simulation parameters, that played an important role in LowModeMD, were found to be Rejection and Iteration Limits. Thus, hundreds of calculations have been run with varying Rejection and Iteration Limits values aiming to identify a set of these parameters, resulting into CCS values in a good agreement with the experimental data. Such parameters were agreed to be equal to 50000 cycles. Further testing these values has proved the argument that they were the cause of the discrepancy, and, moreover, showed the importance of running at least a dozen of protocol instances in order to build a better statistics out of the resulting CCS values. This is due to the fact that even though the conformational search with the new parameters has performed much better than previously, LowModeMD is still a stochastic algorithm, thus, one can always end up by chance with results that have not completely explored the conformational space.

After verifying the reliability of the protocol with the new simulation parameters, in the second part of the project, a focus was shifted to performing Collision Dynamics Simulations. Two isomers were chosen for the analysis — Naringenin-4'-O-β-D-Glucuronide and Naringenin-7-

O-β-D-glucuronide with the aim of modelling their collision processes with the background gas and reconstructing corresponding experimental Mass Spectrometry (MS) spectra. The latter ones potentially could be used to theoretically identify the compounds. Thorough benchmarking has led to identification of the most appropriate input parameters for VENUS96 calculations and thousands of trajectories were run to obtain statistics. Working collision energy was defined to be $E_{coll} = 400^{kcal}/_{mol}$ as it was close to the experimental value and, importantly, provided sufficient percentage of reactive trajectories. Obtained reactive trajectories served as a basis for fragmentation analysis, performed with the help of *bash* scripts. The results reveal that indeed it is possible to distinguish the isomers by comparing their theoretical MS spectra thanks to several distinctive picks. In addition to this, simulations showed that, while having a big set of identical fragments, Naringenin-4'-O-β-D-Glucuronide occasionally experiences different to Naringenin-7-O-β-D-glucuronide channels of fragmentation, what makes it feasible to detect a difference. Moreover, a lower percentage of reactive trajectories was observed for Naringenin-4'-O-β-D-Glucuronide, therefore prompting more calculations run with this isomer to build appropriate statistics.

References

[1] Takashi Iyanagi. “Molecular Mechanism of Phase I and Phase II Drug-Metabolizing Enzymes: Implications for Detoxification”. In: *Int. Rev. Cytol.* 260 (2007), pp. 35–112. ISSN: 00747696. DOI: 10.1016/S0074-7696(06)60002-8.

[2] J. Caldwell, I. Gardner, and N. Swales. “An introduction to drug disposition: The basic principles of absorption, distribution, metabolism, and excretion”. In: *Toxicol. Pathol.* Vol. 23. 2. 1995, pp. 102–114. ISBN: 0192-6233. DOI: 10.1177/019262339502300202.

[3] Gordon J. Dear, Claire Beaumont, Andrew Roberts, Bianca Squillaci, Steve Thomas, Mike Nash, and Donna Fraser. “Approaches for the rapid identification of drug metabolites in early clinical studies”. In: *Bioanalysis* 3.2 (2011), pp. 197–213. ISSN: 17576180. DOI: 10.4155/bio.10.186.

[4] Claire Beaumont, Graeme C. Young, Tom Cavalier, and Malcolm A. Young. “Human absorption, distribution, metabolism and excretion properties of drug molecules: A plethora of approaches”. In: *Br. J. Clin. Pharmacol.* 78.6 (2014), pp. 1185–1200. ISSN: 13652125. DOI: 10.1111/bcp.12468.

[5] Sofia Moco, Jacques Vervoort, Sofia Moco, Raoul J. Bino, Ric C.H. De Vos, and Raoul Bino. “Metabolomics technologies and metabolite identification”. In: *TrAC - Trends Anal. Chem.* (2007). ISSN: 01659936. DOI: 10.1016/j.trac.2007.08.003.

[6] G.J. Dear, Andrew D. Roberts, Claire Beaumont, S.E. North, and J. Chromatogr. “New Horizons in Predictive Drug Metabolism and Pharmacokinetics”. In: *Anal. Technol. Biomed. Life Sci.* 2 (2008), pp. 182–190.

[7] Kerem Bingol, Lei Bruschweiler-Li, Cao Yu, Arpad Somogyi, Fengli Zhang, and Rafael Brüschweiler. “Metabolomics beyond Spectroscopic Databases: A Combined MS/NMR Strategy for the Rapid Identification of New Metabolites in Complex Mixtures”. In: *Anal. Chem.* 87.7 (2015), pp. 3864–3870. ISSN: 15206882. DOI: 10.1021/ac504633z. arXiv: 15334406.

[8] Gordon J. Dear, Jordi Munoz-Muriedas, Claire Beaumont, Andrew Roberts, Jayne Kirk, Jonathan P. Williams, and Iain Campuzano. “Sites of metabolic substitution: Investigating metabolite structures utilising ion mobility and molecular modelling”. In: *Rapid Commun. Mass Spectrom.* 24.21 (2010), pp. 3157–3162. ISSN: 09514198. DOI: 10.1002/rcm.4742. arXiv: NIHMS150003.

[9] Iain Campuzano, Matthew F. Bush, Carol V. Robinson, Claire Beaumont, Keith Richardson, Hyungjun Kim, and Hugh I. Kim. “Structural characterization of drug-like compounds by ion mobility mass spectrometry: Comparison of theoretical and experimentally derived nitrogen collision cross sections”. In: *Anal. Chem.* 84.2 (2012), pp. 1026–1033. ISSN: 00032700. DOI: 10.1021/ac202625t.

[10] Cris Lapthorn, Frank S. Pullen, Babur Z. Chowdhry, Patricia Wright, George L. Perkins, and Yanira Heredia. “How useful is molecular modelling in combination with ion mobility mass spectrometry for ’small molecule’ ion mobility collision cross-sections?” In: *Analyst* 140.20 (2015), pp. 6814–6823. ISSN: 13645528. DOI: 10.1039/c5an00411j.

[11] Eamonn Reading, Jordi Munoz-Muriedas, Andrew D. Roberts, Gordon J. Dear, Carol V. Robinson, and Claire Beaumont. “Elucidation of Drug Metabolite Structural Isomers Using Molecular Modeling Coupled with Ion Mobility Mass Spectrometry”. In: *Anal. Chem.* 88.4 (Feb. 2016), pp. 2273–2280. ISSN: 0003-2700. DOI: 10.1021/acs.analchem.5b04068. URL: http://pubs.acs.org/doi/10.1021/acs.analchem.5b04068.

[12] Jianbo Liu, Kihyung Song, William L. Hase, and Scott L. Anderson. “Direct dynamics study of energy transfer and collision-induced dissociation: Effects of impact energy, geometry, and reactant vibrational mode in H2CO+- Ne collisions”. In: *J. Chem. Phys.* 119.6 (2003), pp. 3040–3050. ISSN: 00219606. DOI: 10.1063/1.1588634.

[13] Christopher Becker, Francisco A Fernandez-Lima, and David H Russell. “Ion mobility-mass spectrometry: a tool for characterizing the petroleum.” In: *Spectrosc. (Duluth, MN, United States)* 24.4 (2009), pp. 38–42. ISSN: 0887-6703.

[14] Zhiyu Li, Stephen J. Valentine, and David E. Clemmer. “Complexation of amino compounds by 18C6 improves selectivity by IMS-IMS-MS: Application to petroleum characterization”. In: *J. Am. Soc. Mass Spectrom.* 22.5 (2011), pp. 817–827. ISSN: 10440305. DOI: 10.1007/s13361-011-0105-0.

[15] Frances D L Kondrat, Weston B. Struwe, and Justin L P Benesch. “Native mass spectrometry: Towards high-throughput structural proteomics”. In: *Struct. Proteomics High-Throughput Methods Second Ed.* 2014, pp. 349–371. ISBN: 9781493922307. DOI: 10.1007/978-1-4939-2230-7_18.

[16] Brandon T. Ruotolo, Justin L.P. Benesch, Alan M. Sandercock, Suk Joon Hyung, and Carol V. Robinson. “Ion mobility-mass spectrometry analysis of large protein complexes”. In: *Nat. Protoc.* 3.7 (2008), pp. 1139–1152. ISSN: 17542189. DOI: 10.1038/nprot.2008.78.

[17] Izhak Michaelevski, Noam Kirshenbaum, and Michal Sharon. “T-wave Ion Mobility-mass Spectrometry: Basic Experimental Procedures for Protein Complex Analysis”. In: *J. Vis. Exp.* 41 (2010). ISSN: 1940-087X. DOI: 10.3791/1985. URL: http://www.jove.com/index/Details.stp?ID=1985.

[18] Waters Corporation. *Waters SYNAPT G2 High Definition Mass Spectrometry*. 2009. URL: http://blog.waters.com/tof-hrms-rescuing-low-exposure-bioanalysis-applications-from-chemical-noise (visited on 01/12/2019).

[19] Matthias Mann, Chin Kai Meng, and John B. Fenn. "Interpreting Mass Spectra of Multiply Charged Ions". In: *Anal. Chem.* 61.15 (1989), pp. 1702–1708. ISSN: 15206882. DOI: 10.1021/ac00190a023.

[20] John B. Fenn. "Electrospray ionization mass spectrometry: How it all began". In: *J. Biomol. Tech.* 13.3 (2002), pp. 101–118. ISSN: 15240215. DOI: 10.1126/science.2675315. arXiv: PMC2279858.

[21] W. Corporation. *What Types of Instruments Are Used?* URL: https://www.waters.com/waters/en%7B%5C_%7DGB/What-Types-of-Instruments-Are-Used%7B%5C%%7D3F/nav.htm?locale=en%7B%5C_%7DGB%7B%5C&%7Dcid=10090937.

[22] Matthew F. Bush, Zoe Hall, Kevin Giles, John Hoyes, Carol V. Robinson, and Brandon T. Ruotolo. "Collision cross sections of proteins and their complexes: A calibration framework and database for gas-phase structural biology". In: *Anal. Chem.* 82.22 (2010), pp. 9557–9565. ISSN: 00032700. DOI: 10.1021/ac1022953.

[23] Valentina D'Atri, Massimiliano Porrini, Frédéric Rosu, and Valérie Gabelica. "Linking molecular models with ion mobility experiments. Illustration with a rigid nucleic acid structure". In: *J. Mass Spectrom.* 50.5 (2015), pp. 711–726. ISSN: 10969888. DOI: 10.1002/jms.3590. arXiv: NIHMS150003.

[24] Samuel I. Merenbloom, Tawnya G. Flick, and Evan R. Williams. "How hot are your ions in TWAVE ion mobility spectrometry?" In: *J. Am. Soc. Mass Spectrom.* 23.3 (2012), pp. 553–562. ISSN: 10440305. DOI: 10.1007/s13361-011-0313-7. arXiv: NIHMS150003.

[25] Yueyang Zhong, Suk Joon Hyung, and Brandon T. Ruotolo. "Characterizing the resolution and accuracy of a second-generation traveling-wave ion mobility separator for biomolecular ions". In: *Analyst* 136.17 (2011), pp. 3534–3541. ISSN: 13645528. DOI: 10.1039/c0an00987c.

[26] Jody C. May et al. "Conformational ordering of biomolecules in the gas phase: Nitrogen collision cross sections measured on a prototype high resolution drift tube ion mobility-mass spectrometer". In: *Anal. Chem.* 86.4 (2014), pp. 2107–2116. ISSN: 00032700. DOI: 10.1021/ac4038448.

[27] Charles C. Kirkpatrick and Larry A. Viehland. "Interaction potentials for the thallium ion-rare gas systems". In: *Chem. Phys.* 120.2 (1988), pp. 235–238. ISSN: 03010104. DOI: 10.1016/0301-0104(88)87169-6.

[28] Christian Bleiholder, Thomas Wyttenbach, and Michael T. Bowers. "A novel projection approximation algorithm for the fast and accurate computation of molecular collision cross sections (I). Method". In: *Int. J. Mass Spectrom.* 308.1 (2011), pp. 1–10. ISSN: 13873806. DOI: 10.1016/j.ijms.2011.06.014.

[29] Edward Mack. “Average cross-sectional areas of molecules by gaseous diffusion methods”. In: *J. Am. Chem. Soc.* 47.10 (1925), pp. 2468–2482. ISSN: 15205126. DOI: 10.1021/ja01687a007.

[30] Alexandre A. Shvartsburg and Martin F. Jarrold. “An exact hard-spheres scattering model for the mobilities of polyatomic ions”. In: *Chem. Phys. Lett.* 261.1-2 (1996), pp. 86–91. ISSN: 00092614. DOI: 10.1016/0009-2614(96)00941-4.

[31] M. F. Mesleh, J. M. Hunter, A. A. Shvartsburg, G. C. Schatz, and M. F. Jarrold. “Structural Information from Ion Mobility Measurements: Effects of the Long-Range Potential”. In: *J. Phys. Chem.* 100.40 (1996), pp. 16082–16086. ISSN: 1089-5639. DOI: 10.1021/jp963709u. arXiv: arXiv:1011.1669v3. URL: http://pubs.acs.org/doi/abs/10.1021/jp963709u.

[32] Alexandre A. Shvartsburg, Bei Liu, K. W. Michael Siu, and Kai Ming Ho. “Evaluation of ionic mobilities by coupling the scattering on atoms and on electron density”. In: *J. Phys. Chem. A* 104.26 (2000), pp. 6152–6157. ISSN: 10895639. DOI: 10.1021/jp0004765.

[33] Ken Shoemake. “Uniform Random Rotations”. In: *Graph. Gems III (IBM Version)*. 1992, pp. 124–132. ISBN: 0124096735 and 0124096700 and 0126683204. DOI: 10.1016/B978-0-08-050755-2.50036-1. URL: http://linkinghub.elsevier.com/retrieve/pii/B9780080507552500361.

[34] Erik G. Marklund, Matteo T. Degiacomi, Carol V. Robinson, Andrew J. Baldwin, and Justin L.P. Benesch. “Collision cross sections for structural proteomics”. In: *Structure* 23.4 (2015), pp. 791–799. ISSN: 18784186. DOI: 10.1016/j.str.2015.02.010.

[35] Vaibhav Shrivastav, Minal Nahin, Christopher J. Hogan, and Carlos Larriba-Andaluz. “Benchmark Comparison for a Multi-Processing Ion Mobility Calculator in the Free Molecular Regime”. In: *J. Am. Soc. Mass Spectrom.* 28.8 (2017), pp. 1540–1551. ISSN: 18791123. DOI: 10.1007/s13361-017-1661-8.

[36] Tianyang Wu, Joseph Derrick, Minal Nahin, Xi Chen, and Carlos Larriba-Andaluz. “Optimization of long range potential interaction parameters in ion mobility spectrometry”. In: *J. Chem. Phys.* 148.7 (2018). ISSN: 00219606. DOI: 10.1063/1.5016170.

[37] Valérie Gabelica and Erik Marklund. “Fundamentals of ion mobility spectrometry”. In: *Curr. Opin. Chem. Biol.* 42 (2018), pp. 51–59. ISSN: 18790402. DOI: 10.1016/j.cbpa.2017.10.022. arXiv: 1709.02953.

[38] Carlos Larriba and Christopher J. Hogan. “Free molecular collision cross section calculation methods for nanoparticles and complex ions with energy accommodation”. In: *J. Comput. Phys.* 251 (2013), pp. 344–36. ISSN: 00219991. DOI: 10.1016/j.jcp.2013.05.038.

[39] Simon A. Ewing, Micah T. Donor, Jesse W. Wilson, and James S. Prell. “Collidoscope: An Improved Tool for Computing Collisional Cross-Sections with the Trajectory Method”. In: *J. Am. Soc. Mass Spectrom.* 28.4 (2017), pp. 587–596. ISSN: 18791123. DOI: 10.1007/s13361-017-1594-2.

[40] Estefanía Rossich Molina, Daniel Ortiz, Jean Yves Salpin, and Riccardo Spezia. "Elucidating collision induced dissociation products and reaction mechanisms of protonated uracil by coupling chemical dynamics simulations with tandem mass spectrometry experiments". In: *J. Mass Spectrom.* 50.12 (2015), pp. 1340–1351. ISSN: 10969888. DOI: 10.1002/jms.3704.

[41] Paul M. Mayer and Clement Poon. "The mechanisms of collisional activation of ions in mass spectrometry". In: *Mass Spectrom. Rev.* 28.4 (2009), pp. 608–639. ISSN: 02777037. DOI: 10.1002/mas.20225. arXiv: NIHMS150003.

[42] Lekha Sleno, Dietrich A. Volmer, Borislav Kovacević, and Zvonimir B. Maksić. "Gas-phase dissociation reactions of protonated saxitoxin and neosaxitoxin". In: *J. Am. Soc. Mass Spectrom.* 15.4 (2004), pp. 462–477. ISSN: 10440305. DOI: 10.1016/j.jasms.2003.11.013.

[43] Julia Laskin and Jean H. Futrell. "Surface-induced dissociation of peptide ions: Kinetics and dynamics". In: *J. Am. Soc. Mass Spectrom.* 14.12 (2003), pp. 1340–1347. ISSN: 10440305. DOI: 10.1016/j.jasms.2003.08.004.

[44] Pascal De Sainte Claire and William L. Hase. "Thresholds for the collision-induced dissociation of clusters by rare gas impact". In: *J. Phys. Chem.* 100.20 (1996), pp. 8190–8196. ISSN: 00223654. DOI: 10.1021/jp953622t.

[45] Pascal De Sainte Claire, Gilles H. Peslherbe, and William L. Hase. "Energy transfer dynamics in the collision-induced dissociation of Al6and Al13clusters". In: *J. Phys. Chem.* 99.20 (1995), pp. 8147–8161. ISSN: 00223654. DOI: 10.1021/j100020a043.

[46] W. L. Hase et al. "VENUS96: A General Chemical Dynamics Computer Program". In: *Quantum Chem. Progr. Exch. Bull.* 16 (1996), p. 671. ISSN: 09406808. DOI: 10.1016/0022-2364(83)90059-8.

[47] Emilio Martínez-Nuez, Antonio Fernández-Ramos, Saulo A. Vázquez, Jorge M.C. Marques, Mingying Xue, and William L. Hase. "Quasiclassical dynamics simulation of the collision-induced dissociation of Cr (CO)6+with Xe". In: *J. Chem. Phys.* 123.15 (2005). ISSN: 00219606. DOI: 10.1063/1.2044687.

[48] Zahra Homayoon, Subha Pratihar, Edward Dratz, Ross Snider, Riccardo Spezia, George L. Barnes, Veronica Macaluso, Ana Martin Somer, and William L. Hase. "Model Simulations of the Thermal Dissociation of the TIK(H+)2Tripeptide: Mechanisms and Kinetic Parameters". In: *J. Phys. Chem. A* 120.42 (2016), pp. 8211–8227. ISSN: 15205215. DOI: 10.1021/acs.jpca.6b05884.

[49] Oussama Meroueh and William L. Hase. "Collisional activation of small peptides". In: *J. Phys. Chem. A* 103.20 (1999), pp. 3981–3990. ISSN: 10895639. DOI: 10.1021/jp984712b.

[50] Jianbo Liu, Brady W. Uselman, Jason M. Boyle, and Scott L. Anderson. "The effects of collision energy, vibrational mode, and vibrational angular momentum on energy transfer and dissociation in NO 2 +-rare gas collisions: An experimental and trajectory study". In: *J. Chem. Phys.* 125.13 (2006). ISSN: 00219606. DOI: 10.1063/1.2229207.

[51] William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters". In: *J. Chem. Phys.* 76.1 (1982), pp. 637–649. ISSN: 00219606. DOI: 10.1063/1.442716. arXiv: arXiv:1011.1669v3.

[52] Xiche Hu, William L. Hase, and Tony Pirraglia. "Vectorization of the general Monte Carlo classical trajectory program VENUS". In: *J. Comput. Chem.* 12.8 (1991), pp. 1014–1024. ISSN: 1096987X. DOI: 10.1002/jcc.540120814.

[53] James J.P. Stewart. "MOPAC: A semiempirical molecular orbital program". In: *J. Comput. Aided. Mol. Des.* 4.1 (1990), pp. 1–103. ISSN: 0920654X. DOI: 10.1007/BF00128336. arXiv: 916061.

[54] James J.P. Stewart. "Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements". In: *J. Mol. Model.* 13.12 (2007), pp. 1173–1213. ISSN: 16102940. DOI: 10.1007/s00894-007-0233-4.

[55] David Weininger. "SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules". In: *J. Chem. Inf. Comput. Sci.* 28.1 (1988), pp. 31–36. ISSN: 00952338. DOI: 10.1021/ci00057a005.

[56] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. "Comparison of multiple amber force fields and development of improved protein backbone parameters". In: *Proteins Struct. Funct. Genet.* 65.3 (2006), pp. 712–725. ISSN: 08873585. DOI: 10.1002/prot.21123. arXiv: 0605018 [q-bio].

[57] David A. Case, Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. "The Amber biomolecular simulation programs". In: *J. Comput. Chem.* 26.16 (2005), pp. 1668–1688. ISSN: 15213773. DOI: 10.1002/anie.198403641. arXiv: NIHMS150003.

[58] Araz Jakalian, Bruce L. Bush, David B. Jack, and Christopher I. Bayly. "Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method". In: *J. Comput. Chem.* 21.2 (2000), pp. 132–146. ISSN: 01928651. DOI: 10.1002/(SICI)1096-987X(20000130)21:2<132::AID-JCC5>3.0.CO;2-P. arXiv: arXiv:1011.1669v3.

[59] Araz Jakalian, David B. Jack, and Christopher I. Bayly. "Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation". In: *J. Comput. Chem.* 23.16 (2002), pp. 1623–1641. ISSN: 01928651. DOI: 10.1002/jcc.10128. arXiv: arXiv:1011.1669v3.

[60] Arthur Dalby, James G. Nourse, W. Douglas Hounshell, Ann K.I. Gushurst, David L. Grier, Burton A. Leland, and John Laufer. "Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited". In: *J. Chem. Inf. Comput. Sci.* 32.3 (1992), pp. 244–255. ISSN: 00952338. DOI: 10.1021/ci00007a012.

[61] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. "Open Babel: An Open chemical toolbox". In: *J. Cheminform.* 3.10 (2011). ISSN: 17582946. DOI: 10.1186/1758-2946-3-33.

[62] OpenBabel. *The Open Babel Package*. 2013. URL: http://openbabel.org.

[63] M. J. Frisch et al. *Gaussian 09, Revision A.02*. Wallingford CT, 2016. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

[64] Paul Labute. "LowModeMD - Implicit low-mode velocity filtering applied to conformational search of macrocycles and protein loops". In: *J. Chem. Inf. Model.* 50.5 (2010), pp. 792–800. ISSN: 15499596. DOI: 10.1021/ci900508k.

[65] Martin Santavy and P. Labute. "SVL: The Scientific Vector Language". In: *Chem. Comput. Gr. Inc.* (2000). URL: https://www.chemcomp.com/journal/svl.htm.

[66] B. Iglewicz and D. Hoaglin. "How to Detect and Handle Outliers". In: *ASQC Basic Ref. Qual. Control Stat. Tech.* 16 (1993).

[67] Junmei Wang, Wei Wang, Peter A. Kollman, and David A. Case. "Automatic atom type and bond type perception in molecular mechanical calculations". In: *J. Mol. Graph. Model.* 25.2 (2006), pp. 247–260. ISSN: 10933263. DOI: 10.1016/j.jmgm.2005.12.005.

[68] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. "Development and testing of a general Amber force field". In: *J. Comput. Chem.* 25.9 (2004), pp. 1157–1174. ISSN: 01928651. DOI: 10.1002/jcc.20035. arXiv: z0024.